BARCH: A Business Analytics Problem Formulation and Solving Framework

CONTEXT STATEMENT SUBMITTED TO MIDDLESEX UNIVERSITY

FOR A DOCTORATE IN PROFESSIONAL STUDIES

by

PUBLIC WORKS

Murphy Choy

INSTITUTE FOR WORK BASED LEARNING

MIDDLESEX UNIVERSITY LONDON

May 2015

Contents

Summary

The BARCH framework is a business framework that is specifically formulated to help analysts and management who want to identify and formulate a scenario to which Analytics can be applied and the outcome will have a direct impact on the business. This is the overarching public work that I have used extensively in various projects and research. This framework has been developed initially in the banking sector and has evolved progressively with successive projects. The framework's name represents five aspects for the formulation and identification of an area that one can use Analytics to answer. The five aspects are Business, Analytics, Revenue, Cost and Human. The five aspects represent the entire system and approach to the identification, formulation, understanding and modelling of Analytic problems. The five aspects are not necessarily sequential but are interrelated in some ways where certain aspects are dependent on the other aspects. For example, revenue and cost are related to business and depend on the business from which they are derived.

However, in most practices involving Analytics, Analytics are conducted independent of business and the techniques in Analytics are not derived from business directly. This lack of harmony between business and Analytics creates an unfortunate combination of factors that has led to the failure of Analytics projects for many businesses. In intensely practising Analytics and critically reflecting on every piece of work I have done, I have learned the importance of combining knowledge with skills and experience to come up with new knowledge and a form of practical wisdom. I also realize now the importance of understanding fields that are not directly related to my field of specialization. Through this context statement I have been able to increase the articulation of my thinking and the complexities of practice through approaches to knowledge such as transdisciplinarity which further supports the translation of what I can do and what needs to be done in a way that business clients can understand. Having the opportunity to explore concepts new to me from other academic fields and seeking their relevance and application in my own area of expertise has helped me considerably in the ongoing development of the BARCH framework and successful implementation of Analytics projects.

I have selected the results of three projects published in papers that are listed in Appendices A-C to demonstrate how the model can be applied to solve problems successfully compared to other frameworks. The evolution of the model involves a continual feedback loop of learning from each successive project which contributes to the BARCH model being able to not only continuously demonstrate its applicability to various problems but to consistently produce better and more refined results.

The majority of analytical models applied to the many problems in the business environment address the problems only superficially (Bose, 2009; Krioukov et. Al., 2011), that is without understanding the impact on the business as a whole. Many Analytics projects have not delivered the promised impact because the models applied are overly complicated (Stubbs, 2013) to solve the root causes of the business problem. This situation is compounded by an increasing number of analysts applying Analytics to business problems without a proper understanding of the context,

technique and environment (Stubbs, 2013). While many experts in the field interpret the problem as a multidisciplinary problem, the problem is in my opinion transdisciplinary in nature.

Context Statement

1. Introduction

The preliminary section of this document is concerned with my research journey to achieve what I have so far, and the formative personal and professional influences that have indicated the directions in which I have developed the works and myself. This is followed by a section dedicated to a detailed examination of the significance of the professional context in and for which the works have been developed and in particular the role of an Analytics Expert. The potential impact of the professional context on Analytics, Experts' opinion and analytical activity in general will be presented and positioned in relevant professional and scientific literature. At the end of each public work, I will summarize the learning from each case and problem through a critique of the failings of the current models, the reason for their failures and how the new model works. I will explore my learning process and how that led to improvement and refinement of the framework. In the process, I will also examine a range of models in use to highlight the differences with reference to my model and to better articulate the complexity of the field of Business Analytics. I will also use the opportunity to describe the issues at hand with regards to Business Analytics.

I will then look at each of the works through the case studies and see how they have moved me towards the development of a framework in many industries. I believe that the on-going development of the framework will be best served by this exploration into its development so far. The development of the model has been achieved through the accumulation of knowledge, insights and wisdom derived through experience. The core essence of the model is the result of careful distillation of the experience I have had with each project. I will highlight how my approach shaped by my professional training and fused with my considerable practice experience, which in turn influenced my repertoire of methodologies, have enabled me to implement Analytics successfully. I will also evaluate how the new framework compares against other existing frameworks through their application in the case studies, which will deepen my own critique of my framework, and the learning that have led to its development.

1.1 Expertise and Formulation of BARCH

What I do as a professional in Analytics involves knowledge (not necessarily expertise) of several disciplines. However the challenge is describing what that means and how working with several disciplines can be coherent and comprehensive which it has to be if it is to contribute to knowledge for the future. There are a number of conceptualisations of working with other disciplines.

Interdisciplinarity is when fields that are related to each other come together to solve a problem like nursing and medical devices training. Since these are already linked in some way and in the context of work environments it is understandable that there will be certain issues arising that need the input of both. Another example is the combination of anthropology and epidemiology working hand in hand to understand the spread of the Ebola virus. The fields don't share the same relationship as in the previous example, but demonstrate that so long as there is common ground, they can intertwine and develop interesting results. The approach was critical in uncovering the nature of the virus and how it crosses the human barrier and spreads due to cultural practices which involve contact with bats that act as reservoirs for the viruses. In the case of Interdisciplinarity, there is always the aspect of a common basis for the disciplines to work with one another (Hammer and Soderqvist, 2000; Ramadier, 2004). Multidisciplinarity is an approach to studying a research topic or problem in one discipline with contributions from other disciplines that bring different perspectives of the disciplines and usually enrich the understanding of the problem and hopefully arrive at a solution (Balsiger, 2004). The approach can positively contribute to knowledge of the home discipline but it usually does not contribute to the other disciplines nor alter their paradigms or methodologies. An example might be nursing and management and economics (Gibbon, 2001).

However, in the above examples, the practice does not necessarily produce anything that fundamentally challenges or changes the paradigms or approaches of the disciplines involved in the practice of Analytics, it is often the case that people will adopt either approach to solve their problems. The key pitfall for both approaches is that they only attempt to pull information and methodologies together in a superficial way to solve a problem or to integrate methodology from one discipline into another to solve the problem. Such approaches are unlikely to be holistic. The Ebola case provides an interesting case where anthropological studies provide insights into the medical practice on the origins and spread of Ebola. Despite this link, there appears to have been no attempt to derive medical insights combining with those of anthropology which might yield interesting approaches to containment of dangerous diseases generally.

I would describe myself as a transdisciplinary practitioner and have come to believe that transdisciplinarity best describes what business Analytics practitioners do and how they think in their work. Transdisciplinarity as defined by Nicolescu (2005) concerns problems that are between disciplines, across different disciplines, or beyond all constraints of discipline. Analytics is something that is not constrained by a specific discipline or several disciplines. Disciplines work together with the intention of being changed by the encounter. Analytics has been applied to a wide range of problems that cross multiple areas in businesses, scientific studies and even social research. The approach is the understanding of the present world and the search is for coherence rather than unity, for approaching knowledge as that which is constantly evolving rather than bound in a body of knowledge.

Transdisciplinarity is characterized by the refusal of formulating any methodology that is constrained by a single core discipline and concentrating on joint problem solving through the science-technology-society tripartite approach. An example is the application of Analytics to fight social ills such as petty crimes (Casey, 2013). It is coming at a problem with no firmly held paradigms and being open to solutions for the benefit of as many stakeholders as possible and also to paradigmatic and methodological change in one's own discipline. Analytics fits very well into transdicsiplinarity because by its very nature it is non paradigmatic, it is open to all kinds of data input and all kinds of

knowledge. Every Analytics problem has multiple facets and requires deep understanding in each facet. All Analytics problems are real world problems and they are not theorized or conceptual problems which naturally places them right at the core of transdisciplinary practices (Gibbon et. Al., 1994; Lawrence and Despre, 2004). The models are developed and shaped by the quality and variety of the data made available as inputs and is in a constant state of evolution and revision. For example, a nurse-scheduling problem is a management problem but to solve such a problem, it requires knowledge from other disciplines such as biology, change management, economics, sociology, and/or psychology. The constraints are always changing due to the context (policy making and populace opinions) and the environment (illnesses and epidemics). The requirements from the individual disciplines go beyond the superficial understanding and integrations. The various constraints in the problem demonstrate the complexities in the problems that Analytics faces and why it is important to have a transdisciplinary approach. This is an approach that has given rise to techniques such as Genetic algorithms with its origins in the area of Biology and Genetics or the contribution of physics that inspired Simulated Annealing for solving optimization problems. These examples resonate with and support the transdisciplinary thinking and practices which emphasise the fusion of elements from various disciplines into a common framework as the hall mark of transdisciplinarity (Horlick-Jones and Sime, 2004). Given that Analytics needs the input of multiple disciplines, the approach has to be transdisciplinary in nature. The other key factor is context.

One key aspect to a transdisciplinary approach to solving the problem is the importance of context. The context presents the problem in a particular manner that requires certain specifics to be in that aspect. This coincides with the fact that the starting point in a TD approach to problem solving is context (Stokols, 2006).In conjunction with the context, the application of Analytics rejects the traditional generalization and reductionist approaches of disciplinary studies which does not fully explain the inter-connected aspects of the problem.

1.2 The importance of Context

In my practice, I have experienced the difficulties of applying analytical models without understanding the context. The experience was harrowing and led me to believe there was strong need to develop a simple and structured approach to solving business Analytics problems through the use of a framework. One consistent issue is the changing environments and contexts where Analytics reapplied. Every project or problem brings on a new set of environmental conditions with the contextual elements that require adaptation and it makes application of Analytics very difficult if not impossible. The problem is aggravated by the lack of suitable frameworks at the problem solving level to guide analysts and practitioners to handle the problem in a structured manner.

Given the huge impact Analytics has the potential of making on any organization, enabling and empowering business analysts to quickly understand the context of the problem helps to adjust and refine the approach and also in several instances the problem focus. It also demonstrates the value of Analytics to an organisation that is driven by market forces to constantly adapt and update not only their market approaches but the way they think about their own disciplines to meet such field conditions. In addition, in a world enmeshed in global technology, Analytics has the potential to identity and resolve the problems rapidly if applied comprehensively.

There are several aspects of context to consider when critically engaging with these works. As the developer and agent of change I am part of the context too. I have come to realise through doing this engagement with my public works that I am the product of the interactivity of different disciplines myself. I started out just like any beginner and first experienced the difficulties of applying mathematical models to a business context during a project where I had to do customer segmentation. Even after producing very detailed customer segmentation, there were several difficulties in obtaining the buy in from the clients. The main obstacle was the lack of understanding of their business models. It was extremely frustrating due to the amount of effort put into the segmentation. This was made worse by the management who could not translate what they understood about their business to the analysts. The entire project was both frustrating and enlightening to me. It opened my eyes to the need for a coherent framework where management and analysts can understand one another even if they meet for the first time.

This experience generated deep thoughts about the underlying thinking processes that drive my actions and intuition. Drawing on the work of futurists in the area of Galilean Model (Gary, 2008), the thinking process is like looking at the world through the conventional future, counter future and the creative future (more commonly addressed in futures' studies as probable, possible and preferable). It is not possible to deliver sufficient impact on the business through a conventional (probable) approach or a counter (possible) model which is like a first degree model. Since the conventional approach is tested and the counter model too difficult to apply, only a creative solution will be sufficient. To achieve lasting impact, we need to create a second degree model through a creative (preferable) approach that is revolutionary in nature.

This drove me to further experiment with different approaches and identifies interesting insights to deliver that transformation. As I progressed in my career, the drive to transform thinking pushed me to consistently contribute articles to newsletters, conferences and journals. I attended several major conferences in the area of Analytics and presented several papers and findings. I wanted to share knowledge and augment my own by feedback from the participants. After a few years of attendance at these conferences, I became increasingly recognized as an 'expert' of sorts in the area of Analytics. I was happy that people were acknowledging my works and to be recognized by my peers. I was invited to be on panel discussions and to sit on several working committees within the conference groups. Eventually, I became a section chair, which is one of the highest honours for someone in the area of Analytics.

Being considered an expert in the field of Analytics poses some cultural challenges for me. The concept of an expert and an amateur varies from culture to culture. This can cause considerable problems if one happens to straddle between Eastern and Western cultures as I do. Eastern culture refers to the Oriental cultures such as Chinese, Indian, Japanese, Korean, South East Asian and Central Asia which covers much of Asia while Western culture can be viewed primarily as the Latin culture (Berger, 1997).In most cases, Asian and Islamic Nations are considered Eastern Cultures while European origin countries are considered as Western Culture. Literature (Irate, 2002) does not specifically demarcate the geographical locations but separates the worlds by the nature of the culture acknowledging the localization and specialization of the culture in most cases (Mestrovic, 1994). In Eastern culture, the concept of an expert is more akin to a senior person with good experience across multiple areas rather than just a single area of expertise. However in the west, experts are really just professionals who are really proficient and focused in a specific single area of

work (Dreyfus & Dreyfus, 2005; Ericsson, 2000; Ericsson et. Al., 2006).It is important to understand the concept of expertise and to formalize it in a way that can be commonly discussed across cultures without misunderstanding the underlying requirements. At the same time, the diverse and transdisciplinary nature of Analytics makes it difficult to pinpoint expertise. There have been comments that the field is now filled with 'Parrot' experts who are just parroting comments made by the true experts (Smith, 2014).Thus a proper definition as well as good discussion of the notion of expertise is critical.

The inherent nature of technology in modern society inspires cognitive behaviour such as Neomania, a term coined by Rolf Dobelli (Dobelli, 2013) which is partly inspired by the book Antifragile (Taleb, 2012). This unfortunate development in modern society is one of the key drivers to instability of the market. Since the beginning of the modern practice of Analytics in 2007, there have been several iterations of updates and revamps of the original definition. New developments such as big data Analytics and in-memory Analytics has generated a craze in the market that practitioners start branding themselves with various titles and credentials. This dressing of individuals as experts presents a problem for businesses which have difficulty determining the level of expertise. By discussing the nature of expertise as well as the nature of transdisciplinarity, I can attempt to pinpoint the nature of expertise in Analytics and help the industry to establish standards that can be applied to identify true practitioners as opposed to pseudo-practitioners.

Expertise is difficult to measure or observe with the most direct measure being a test that is similar to professional licensing. Moreover, there are many types of expertise, and possible taxonomies of expertise to consider. For pragmatic reasons, I have focused on the areas of expertise most often needed in the field of Analytics that have further assisted me in the discovery of the defining criteria of an expert in Analytics.

One of the most important aspects of expertise is experience. The term 'experience-based experts' has been adopted to describe those whose expertise is recognized not just through the common process of earning or granting of certificates but through their accumulation of years of experience in actual field work or practical experience(Collins and Evans, 2002). Certificates or diplomas are definitely ways to be considered an expert but they do not always serve the function well. Unless they are issued by authorities or through some kind of voting systems, they merely carry the weight that defines a level of proficiency not expertise. Likewise, experience is not the sole defining criterion of expertise. It is a necessary criterion but insufficient. Relying on this sole criterion would result in expertise that is highly irrelevant and comical, such as expertise in drinking water, expertise in sleeping or even expertise in breathing. However, these are not expertises; they are necessary behaviours for survival and are instinctive and biologically driven. There could be other forms of expertise such as the expertise of using a computer mouse which is not instinctive or biologically driven but is a learned necessity and now an automatic skill. A skill can be automatic once it is practised enough like riding a bike. Such 'expertise' usually does not require extensive amounts of practice prior to mastery. However, if one wanted to enter into competition in cycling then mastery would be required. Mastery is usually achieved through learning from a master which is the Eastern notion of apprenticeship (Gu et. Al., 2010). What is learned is not just the skills or craft, the ways of doing things but also mastery in how to be, the values and motivations of doing anything. There are many skills that one could theoretically master by an extensive amount of practice. However it is also the content of the material that will define whether the expertise is legitimate. The content

component is well developed in literature especially in areas of arts and performance arts where the product is easy to interpret and content is critical (Melrose, 2005; Melrose, 2011). One way of describing the process of acquiring expertise is that proposed by Alexander (2003). He suggests three stages: acclimation, competence and expertise.

Acclimation

Acclimation is the initial stage in domain expertise. This stage starts by orienting the learners to a complex, foreign domain. The domain maybe something that the learners have exposure to but not at the level of detail that is common to the learners. For example, most people are acquainted with the basics of arithmetic but they will not have gained the detailed level of knowledge that common arithmetic is a special type of arithmetic called Peano Arithmetic. Likewise, most engineers have a good grasp of Calculus but few will have the knowledge of mathematical analysis of functions. During acclimation, learners acquire limited and fragmented knowledge even though it is possible for the learners to be well versed in specialized topics. Typically, the learner will be exposed to specifics and examples which symbolize and represent the domain in question. The general learners at this stage will not have acquired a cohesive and well-integrated body of domain knowledge (Gelman and Greeno, 1989) as they are exposed to specific examples. The knowledge acquired will not allow acclimating learners to determine the accuracy or relevancy of information (Jetton & Alexander, 1997).

Competence

At the second stage, competence development can be observed by quantitative and qualitative changes in an individual's knowledge base. The individual is exposed to increasing amounts of examples and experience which connects the fragmented knowledge into connected sequences of information and knowledge which complement each other. There must be a clear and sound foundational body of domain knowledge with stronger cohesiveness and principle in structure. The problems in those domains will be more familiar and competent learners will be able to attack such tasks through a combination of surface-level and deep-processing strategies. These knowledge and strategy improvements are the result of an increase in the individual's personal interest in the domain coupled with the increased amount of experience dealing with such problems.

Proficiency/Expertise

Unlike the progression from acclimation to competence, there must be synergistic behaviour among components in order to progress from competence to expertise. The knowledge base of experts must be both broad and deep and they are contributing new knowledge to the domain. The

knowledge base must be coherent and sufficiently connected to form a good body of information. The knowledge is not just synergistic but also multi-faceted to present different perspectives to solve the problem. This multi-faceted aspect also provides various tools and methodologies that have been explored in other areas which might be of use in solving a problem. To generate new knowledge, experts must be well acquainted with the problems and methodologies of the domain and actively engaged in *problem searching.* The experts will generate questions and research into these problems that push the boundaries of the domain. At this stage, the individual motivation of experts is very high which enable maintenance of a high level of engagement over time.

Experts also possess 'cognitive authority' (Merton, 1976). Authority exists as a political concept and while experts are not directly accountable to people unless there are legal reasons, they can exercise some form of authority-like powers over questions of true belief. There are several forms of beliefs and the most common form is the justified true belief form. This form is derived from the works of Plato and is widely discussed (Fine, 2003). So if an expert has a doctorate, and someone believes that the expert is a holder of doctorate, then the person is justified in his beliefs. However, this class of belief has been problematic in the establishment of expertise given the nature of Gettier problems (Timothy, 2007; Gettier, 1963). The Gettier problem is a major problem for experts in Analytics as this group of problems can be encountered in practice regularly. An example will be experts have websites dedicated to their work. Someone believes that an expert has a website dedicated to their work. Therefore the person believes that a person with a website dedicated to his works is an expert. However, this is a false premise. Because the notion of expertise cannot be easily established, this leads to debates and resistance to the notion of expertise.

The nature of cognitive authority allows it to be open to resistance and submission and it cannot be distributed nor be granted. The common view (Turner, 2001) is that the authority has something that others do not have. The approach that experts take with regards to facts or validity of knowledge claims of other experts in the same field is different from the approach which non-experts believe in the experts. The facts in a specialized field are only recognized as facts by people who understand and to do something with them is the expertise of those who are technically trained appropriately in the field. The non-expert is not trained to make sense of the information and they will readily accept the pre-digested views of experts as authoritative (Schmitt, 1932).

Turner (2001) separates expertise into five forms depending on the way they obtain legitimacy from their consumers

Type I is like that of physics, which has gained a kind of universal authority across society in virtue of what everyone believes to be its efficacy. This type of expertise will be deemed to be something that can be applied to all strata of society.

Type II expertise has been granted legitimacy only among a restricted group or sect of adherents. This form of expertise tends to be highly focused on a specialized skill, practice or belief. The common example cited is Theology where its relevance is specific to the practitioner of the religion at a higher level than lay people. Possible examples include Geomancy and other new age movements.

Type III experts, such as a new kind of health or psychological 'therapist', create their own adherents, or groups of followers. This type of expert often introduces new therapies via mass media such as TV shows or paid advertisements. While many of these experts do have professional qualifications such as medical licenses, they often preach in areas unrelated to their original expertise[1].

Type IV and Type V experts have their adherents created for them by professional agencies which set themselves up to promote a new kind of expert, or, like government departments, become specialist consumers of new kinds of expertise. There are professional associations for practitioners in financial sectors such as CFA, ACCA and CAT. Any of these associations are chartered or recognized by government. However, there are other groups set up by practitioners to recognize their own experts. Groups such as SAS user groups or UseR! Groups are practitioner groups.

Turner's research is grounded mainly in the Western world culture. Because of the strong influence of culture on the concept of expertise, it is important to evaluate the fundamental differences in the concept of expertise. Most cross-cultural studies compare and contrast Western analytical thinking with Eastern holistic reasoning. Western analytical reasoning emphasises objects and categories and is driven by formal logic. Eastern holistic reasoning allows for contradictions with a yin–yang [2]view of constant change. This forms a unique difference, which can lead to incompatibility in the reasoning process between cultures. Essentially, the duality of views allows for contradiction so often ignored by practitioners. In the field of Analytics, practitioners are often concerned about the performance of their models in predicting events. This narrow focus is myopic and results in models that are powerful in a narrow and focused way. When generalized, the model fails spectacularly and the results are not often replicated in varying environments. The narrow focus also leads to a fruitless search for the 'Ultimate' model which can solve everything. The recent debates about "Deep Learning" and Watsons [3]have led to much heated discussion about the possibility of such a model. However, both models are too narrowly focused and cannot handle the duality of problems. On the political front, this manifests in the form of diplomatic quarrels between Asian countries and Western countries such as the diplomat arrest case[4] and several other territorial disputes in Asia[5]. In most cases, Western thinkers will search for the root cause while the Eastern thinkers tend to examine the bigger picture (Gries and Peng, 2002).

Expertise in Eastern culture differs slightly from Western in terms of the requirements. Phronesis, otherwise known as practical wisdom, is considered to be the key to expertise in Eastern culture compared to Western cultures. The Chinese concept of an expert is someone who is holistically trained in multiple areas and has distilled a certain level of wisdom from his study subjects. Such distillation process for wisdom involves much practical work and requires an extensive amount of training that allows the person to gather experience. The Western concept of expertise typically considers the in-depth experience of an individual in a particular area with extensive knowledge that

[1] The Dr Oz. Show

[2]Yin – Yang is an oriental concept where yin refers to soft or feminine aspect and yang refers to hard or masculine aspect.

[3]IBM Watsons is a technology developed to challenge the Jeopardy contest. The underlying technology involves a lot of Analytics models and approaches.

[4]Indian Diplomatic Crisis 2014. www.cnn.com/2014/03/14/justice/indian-diplomat-indicted

[5]Hainan Island incident

is far more extensive than lay people. This is similar to the search for the root cause. Thus, so long as an individual has a skill which satisfies the condition of experience and knowledge, the individual will be considered to be an expert. However, knowledge does not necessarily translate into practical wisdom or actual expertise in applying that knowledge in real world situations. At the same time, the concept of an expert in a very specialized field can also result in an expertise that is irrelevant to the real world. An expert in eating competitions will find that his or her skills are irrelevant to the real world as a practical survival skill. So while the individual will be an expert in eating competitions, this expertise might not be viewed in the same way between different cultures. The Eastern concept of expertise requires the individual to be an expert in a way that the skill is relevant to the real world. Thus an eating expert will not likely be viewed as an expert while a horse shoe maker will be. This strong emphasis on practical wisdom is the key difference and has been applied in knowledge transfer in Eastern settings (Gu et. al., 2010).

In Eastern culture, the Type I experts will be considered experts. This recognition derives primarily from the widespread acknowledgement of the efficacy of scientific knowledge universally. Type II experts, depending on the context in Eastern culture, may or may not be considered an expert. The key problem of recognition for Type II expertise derives itself from the close-knit nature of Eastern society (Hofstede, 1993). Eastern culture is generally more trusting of their family member or clan members compared to strangers. Recognition between clans is rare and even less so for social societies or clubs. Type III experts are not considered experts in Eastern culture as they are often associated with quack medicine or bogus experts (Langford, 1999). Interestingly, type IV and V can be considered experts if they are recognized by professional agencies that are recognized by the people as professional bodies. This is due to the nature of these professional agencies being regulated or controlled through legal means. In spite of the type of expertise, the various expertises are supported by the concept of knowledge and Phronesis.

Practical wisdom, Phronesis, is commonly defined as knowledge of the proper ends of life and is classified as 'intellectual virtues' (Eikeland, 2008). Aristotle (Jones, 1975) distinguished phronesis from episteme and techne. Episteme is seen as scientific, universal, invariable, context-independent knowledge, while Techne is considered to be variable, craft knowledge. Phronesis is the intellectual virtue that is concerned with practical judgement and informed by reflection (Jones, 1975). Many professions are plagued with a theory–practice gap in which phronesis offers a bridge (Kinsella and Pitman, 2012). Literature in Phronesis (Kinsella and Pitman, 2012) considers six criteria that develop phronetic judgement in professional practice: pragmatic usefulness, persuasiveness, aesthetic appeal, ethical considerations, transformative potential, and dialogic intersubjectivity. These criteria were also suggested in other literature works not directly related to Phronesis research but in the area of knowledge transfer (Gu et. al., 2010).

There is no universal definition of wisdom, however there are several schools of thought about the concept of wisdom. The different schools of thought can be summarized into the implicit group and the explicit group. Implicit theories are based on the beliefs and mental representations of laypersons about wisdom and wise people (Baltes, Glueck, & Kunzmann, 2002; Kunzmann & Baltes, 2003). This is opposed to the explicit theories that are constructions of theorists and researchers about wisdom (Sternberg, 1998). More importantly, theoretical knowledge is knowledge understood at the intellectual level while wisdom is understood at the experiential level. Knowledge is transformed into wisdom only when the truth of the knowledge is realized. If the truth is understood

intellectually rather than experientially, it remains intellectual knowledge and does not transform into wisdom (Naranjo, 1972).

Literature about professional practices are dominated by medical and nursing practices. Case studies about medical professions developing their professional practices have demonstrated how Phronesis or practical wisdom can be discovered through reflective healthcare practice (Frank, 2012). In the case studies, professional practice or reflection begins with an interruption to the routine and questioning the routine (Dunne, 1993; Frank, 2012; Hibbert, 2012; Higgs, 2012). One critical aspect of professional practice is the presence of uncertainty and recognition of the complexities of professional practice (Higgs, 2012; Schön, 1987). Professional practice also created situations which cannot be solved with high level of uncertainty (Macklin and Whiteford , 2012). Professional practitioners must recognize the application of knowledge in practice is a far cry from theory and the messiness of the application should be accepted and embraced. Literature also recognizes that epistemes are needed together with phronesis for effective practice (Kemmis & Smith, 2008; Macklin and Whiteford, 2012).

In my opinion, an expert is certainly a person who has techne, episteme as well as phronesis. The episteme that the person has must be one that has depth and breadth. This is critical as an expert is not just someone who is multidisciplinary or interdisciplinary but ultimately transdisciplinary. His[6] ability to apply his knowledge has to transcend the borders and limitation of his original training and learning. He has to remove the trappings of the original discipline and elevate his knowledge, skills and techniques to a level that transcends the limits that have been placed on it. The expert has to also test his techne and episteme in real life situations and distil the learnings from the experience to form phronesis.

1.3 Concept of An Expert in Analytics

Using the concepts discussed above, I can piece together the various requirements of an expert in Analytics. The expert definitely needs to have extensive experience in the knowledge and practice of Analytics. Analytics requires in depth knowledge of mathematical and statistical models that are technical skills. At the same time, this knowledge needs to be applied to real world problems in a manner that will be useful to the organization. This application requires Phronesis that exists due to the accumulation of experience applying the knowledge in real world practice. Given that the application of the knowledge is relevant to the organization and that Phronesis can be demonstrated, arguably, the individual has satisfied most of the requirements of an expert with the exception of recognition. In this case, the individual can be considered an expert. In this case I am an expert in Business Analytics.

Being an Expert can be intimidating at the beginning and the authority that comes with it can be seductive. It is important to be aware of one's actions and to maintain quality of work, listening and acting on feedback from peers and continuing professional development from practice and from formal inputs. The constant flux of the Analytics world coupled with the ever growing list of models

[6] The use of the masculine gender is for facilitation not from gender bias

which can be used to solve a variety of problem makes it difficult for anyone to be expert in every aspect of Analytics. A few years ago, if you were able to model using Logistic Regression, you could be considered an expert in Analytics. But in the last few months, anyone who cannot even speak Hadoop, Kafka and Deep Learning is not considered even a candidate to be experienced in the field. Much of these have to do with creative marketing and the ever increasing need to differentiate oneself. However in my field, expertise also requires one to be an able translator as I am not only speaking to my qualified peers, I have to be able to communicate with clients with business problems who are looking to Analytics to solve the problem. There will be no management buy in if they do not understand what I am saying about what Analytics is and what it can do and Analytics will not succeed if I do not understand the total context of the clients environment before exploring what they have defined as the problem. Being able to express the models used clearly and explaining them in the light of the business is demanding. In this respect, developing the ability to conceptualize the problem in a way that is understandable by business and can be solved using Analytics is important in the role.

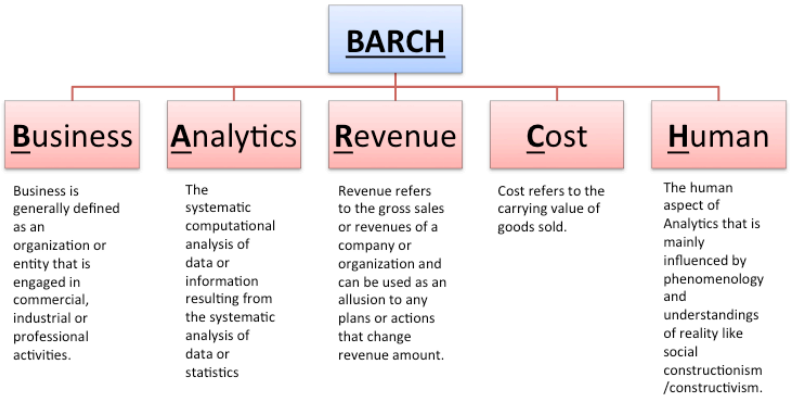
My path to becoming an expert is one that has had several twists to it. I did a diploma in Economics with the London School of Economics under the umbrella of the University of London International Program. The diploma was a life changing experience that shaped much of my outlook in life as well as igniting my interest in applying mathematical models to the real world. Prior to the diploma program, I loved Mathematics as a high school student and had intended to do pure Mathematics when I applied to the University. However, while doing my diploma, I was exposed to Economics modeling which fascinated me endlessly. The topic that stood out was Game theory. Even though the topic was well explained, I realized that I could not apply the model to real world problems. The real world problems were too difficult for me to break down into smaller solvable components. Most real world problems are not isolated problems which do not affect one another, instead they form one giant problem where the solution of one will affect the others. This interrelatedness is the reason for the difficulty of breaking problems down into simpler components. This has parallels in the area of Analytics where optimization cannot be broken into smaller components as the optimum solution from a smaller problem does not add up to the overall optimum solution most of the time. I was challenged to come up with a method to solve real world problems. I entered the University and started pursuing Statistics as my major. The course was rigorous and I learned the basics of Statistical modeling as well as data analysis. Even with this new knowledge, I found myself challenged during my internship at a pharmaceutical company to develop a mathematical model to prove that a particular design can improve the yield of the researchers. During the process, I realized that there was no framework that guided the analyst to do basic modelling works. We just had to try any models that came to our mind and attempt to solve the problem. At this point, I started contemplating the possibility of constructing the bridge that connects the management and the analysts.

The accumulation of my teche, my experience and my approach to knowledge in a complex world have informed how I have constructed that bridge which is not a static structure but one that has increasing capability in adaptation to different contexts.

2. Introduction to the BARCH Framework

BARCH is a framework designed to help analysts and managers working in the area of Business Analytics to identify critical aspects of the business and where Analytics can be applied taking into consideration the human dimensions of business, in order to generate an Analytical model that will deliver a positive impact to the business. The framework's name represents five steps for the formulation and identification of a problem area that one can use Analytics to resolve. The five aspects are Business, Analytics, Revenue, Cost and Human. The five aspects represent the entire process of identification, formulation, understanding and modelling of Analytics problems. The framework emphasizes the importance of business understanding that contributes to the analysts and managers knowledge of the nature of the business. By combining the key company metrics of revenue and cost together with business and human elements, the analysts and managers will be able to understand how the Analytical model that will be built interacts with the other elements and how such data anticipates the implications for each part. This deep understanding of such interplay coupled with an appreciation of the financial metrics will improve the analysts and managers ability to have impactful discussions between the different stakeholders in the business when it comes to using Analytics to solve problems. In the process of these exchanges, the BARCH framework will enhance the ability of the analysts and manager to shape the management's understanding of the 'truth' of their business and to understand how Business Analytics will change the business positively and mitigate it changing the business negatively.

2.1 Components of the BARCH model



Business

Business is generally defined as an organization or entity that is engaged in commercial, industrial or professional activities. The business can be a for-profit entity or a non-profit organization. An individual or a group can undertake the commercial, industrial or professional activity in the business. Businesses have goals or objectives that depend on the nature of business. For example, the for-profit entity will be interested in profit generation. In the case of non-profit entities, they might be defined by other attributes such as number of people helped, lowering literacy rate and

others. The external business environment that businesses operate in also defines the type businesses activities and processes.

Every business operates under a specific business model. The business model explains the process of how an organization creates, delivers, and captures value that can be expressed in economic, social and cultural terms, or other forms of value. In both theory and practice, the business model represents the core aspects of a business, which includes purpose, target audience, products, strategies, infrastructure, organizational structures, business practices, and operational processes and governance. The literature presents very diverse interpretations and definitions of a business model. Literature suggests, through systematic review and analysis of manager responses to a survey, that business models are the product of organizational structures to implement a commercial opportunity (George and Bock, 2011).

Further research in this direction indicates the emphasis on the application of narrative in business model descriptions to be mechanisms from which entrepreneurs can create extraordinarily successful growth firms (George and Bock, 2012). Below is a quote extracted from David Teece's work (2010).

Whenever a business is established, it either explicitly or implicitly employs a particular business model that describes the architecture of the value creation, delivery, and capture mechanisms employed by the business enterprise. The essence of a business model is that it defines the manner by which the business enterprise delivers value to customers, entices customers to pay for value, and converts those payments to profit: it thus reflects management's hypothesis about what customers want, how they want it, and how an enterprise can organize to best meet those needs, get paid for doing so, and make a profit. (Teece, 2010; p.172-194)

Business models are used to describe and classify businesses, but managers within companies also use them to brainstorm for possible future development. Common and popular business models form recipes for creative managers (Baden-Fuller and Morgan, 2010). Business models are also referred to in some instances within the context of accounting for purposes of public reporting.

Most business models are formed from business model frameworks. Business model frameworks represent the core aspect of any company which defines how a company selects its customers defines and differentiates its offerings, defines the tasks it will perform itself and those it will outsource, configures its resource, goes to market, creates utility for customers, and captures profits. Technology centric communities have defined "frameworks" for business modelling. These frameworks attempt to define a rigorous approach to defining business value streams. It is not clear, however, to what extent such frameworks are actually important for business planning. A state of the art review on business model frameworks can be found in Krumeich et al. (2012).

Among the most common and popular business frameworks are:

- Business reference model
- Component business model.

Business reference model

A Business reference model is a model of business that contains the basic (primary) goal or objective of business and serves as a reference for any other purposes. A business reference model describes the business operations of an organization without considering the organizational structures. A business reference model can also depict the relationship between the business processes, business functions, and the business area's business reference model. Most reference models are constructed in layers to offer common basis for the analysis of service components, technology, data, and performance.

Component Business Model

Component Business Model (CBM) was developed by IBM to model and analyse an enterprise. The model creates a logical representation of business components that can be depicted on a single page. The model can be used to analyse the alignment of enterprise strategy with the organization's capabilities and investments, identify any redundant or overlapping business capabilities, analyse sourcing options for the different components (buy or build), prioritizing transformation options and can be used to create a unified roadmap after mergers or acquisitions. The model is usually organized as business components along columns and "operational levels" along rows. The Business components are defined partly as large business areas with characteristic skills, IT capabilities and process. The large business areas usually correspond to key areas such as marketing, finance and human resource. The components dictate into three operational levels are "Direct", "Control" and "Execute" - they separate strategic decisions (Direct), management checks (Control), and business actions (Execute) on business competencies.

Business as a part of BARCH

In the BARCH model, the Business component basically breaks down the business into its constituent components. The Business aspect is focused on the process that drives the business and the objective that it serves. For example, an airport's business is to serve as the port to air traffic or transport while the business of a hospital is to serve the people with proper medical care. It is extremely important to identify the Business and Objective, as they are the main drivers for all aspects of business. It is important to note that business practices are driven by a combination of internal and external factors. The BARCH model does not distinguish the internal and external factors as they are combined together in practice and distinguishing them might not always be possible. This view is also echoed in the literature that will be discussed towards the end of this section.

In my experience, identifying the business is perhaps the most important skill required and task that needs to be done. Even though every business has its objective and business processes, my experience in the field as a practitioner tells me otherwise. The airport case in the case study section is a prime example. In that particular case, even though the business of the department is allocation of service counters to process the passport and air tickets of the departing customers of different airlines, it has to take into consideration the business of the airport to allow as many aircraft to land and depart as possible. However, the business operators and users were not even concerned about the need to get more airlines to fly into the airport. Instead, they were fretting about the need to re-negotiate with the different airlines about their individual arrangements. This is very common in big organizations where business operations have deviated from the original business function significantly.

The fixation on the negotiation initially rendered the project untenable and progress was slow. The operators' working position made it impossible to apply any models to come up with a solution. Despite this, I applied the Business aspect of BARCH and directly resolved the problem once it was clear that the business model was not the same as the operators believed it to be. The understanding of the Business will help any analyst to quickly identify the source of the problem and apply the relevant model.

My works in the realm of hospitals also suggested a lack of understanding of the business process. The key resources in hospitals are beds, medicine, equipment, doctors and nurses. The lack of any components will have a disastrous effect on the operation of the hospital. This motivates the hospitals to be good at scheduling resources especially on the nursing resources aspect as it is heavily regulated. At the same time, nurses are human and have preferences. Thus a proper nurse scheduling is critical in ensuring the quality of care for patients while managing the needs of the nurses. To achieve that, methods incorporating human preferences and the legal requirements are critical.

However, this is unfortunately not the case for several hospitals that I have visited. A good analyst is a trained and detailed observer as the quality of data directly impacts on the success of outcome. My own chronic illnesses such as arthritis, gastroenteritis and infections have necessitated several surgical operations and hospital stays in the past few years. This gave me the opportunity to understand the operations of the business that would otherwise not be possible especially since the people who are hospitalized are not trained to study such operations.

What is really shocking is the lack of understanding of the operations of the hospital by the management despite the management being formed from a selection of staff from the various areas. The management is keen to manage the waiting time of patients and management of beds while the nurses' needs are made subordinate to the former two. This has severe repercussions on the nurses' health both emotionally and physically. I was, on several occasions, able to identify nurses who are given shifts which are barely 8 hours apart impacting on their own health and care for patients.

During a particular stay, I noticed peculiarities in the ward where I was staying that defies the soft constraints of nurse scheduling. Several nurses were reporting in at hours, which certainly did not please them. For a particular nurse, I noticed that she was doing a night shift followed by an afternoon shift that is less than 8 hours apart. While technically legal, I was concerned about her

welfare, as she looked visibly tired. I questioned her schedule and she was glad that someone had noticed that the arrangement was unsound both from a personal and professional perspective. A tired nurse is not an effective one. I chatted with the various matrons to learn about their difficulties.

In this case, I believed the arrangement was not arbitrary but due to the compliance requirements. Through personal experience of an area in which I am not an expert but in fact a customer, my belief was confirmed that Business Analytics can make life better for everyone else but not when it is applied without good understanding of the business.

During my tenure in the banking industry, I faced similar problems where the majority of analysts who deal with Analytics to resolve problems have very little understanding of the operations and processes of the banks. The implementation of credit risk scorecards and Basel II standards exemplifies the problem between the modelers and business. The Basel II standard imposes a standard definition across all the various credit instruments. Because of the nature of the definition, it causes a significant number of problems as the operations and nature of the instruments defy the definition making it difficult to model. The fluid nature of credit instruments also renders the definitions hard to implement or model. One good example is the bad definition in the credit portfolio. The standard bad definition in banks using the Basel II standard is 90+ Days past Due. However, this definition is not always something that the operations people are able to handle especially in cases like mortgages where a 90+ Days past Due implies a near default level which leaves the bank with little time to react. Coming up with a bad definition is also difficult as it is by nature a chicken and egg problem. Thus it takes a considerable amount of understanding of the business to come up with a reasonable definition. This is however problematic when we have to consider the customized definition with the standard Basel II definition. This is where Analytics comes into play and the careful use of Analytics can help to solve the business problems.

One critical factor that will affect the use of BARCH models is the ability to understand businesses and Business Models. There are various types of business models which cater to different types of businesses. There have been substantial studies on the nature of business models, the ontology of business models and the various components of business models (Hedman and Kalling, 2003; Osterwalder, 2004). However, to understand each type of business requires considerable immersion in literature and field work which is a challenging task even for those engaged in the discipline (Osterwalder, Pigneur and Tucci, 2005). However, general principles can be drawn.

The common theme in business models is the value proposition. This is the underlying driving force behind the business and the entire process revolves around it (Johnson, Christensen and Kagermann, 2008). In some literature, business development and evolution (Casadesus-Masanell and Ricart, 2010) consist of two stages, the business model selection stage and the business strategizing stage. A business model is defined as "the logic of the firm, the way it operates and how it creates value for its stakeholders." (Baden-Fuller et. Al., 2010). The business model consists of various processes that are designed to make use of resources in order to achieve the value proposition. The business model is also influenced or in some cases constrained by external conditions such as legal requirements and geographical locations. Once the processes are in place, the management is required to make decisions on the choices of resources and processes that will lead the business to its eventual success or profit. To ensure the continuity of the business to

generate profit, there is a profit formula that determines the key objective of the business that needs to be optimized by the business in order to be successful. This combination of value proposition, profit formula and processes optimization ensure that any business model developed will have incorporated the influence from both internal and external sources.

It is critical in the process of understanding the value proposition to view it neutrally without undue concerns about the success that has been generated in the past as a result of the model. This is because the model might have been performing very well in the past but has diminished through the years and requires a thorough rethink and reinvention. An example would be Nokia who once dominated the market only to be bought over by Microsoft after losing almost 90% of the share due to an unchanging culture and business model. The value proposition might no longer reflect the organization or business. This is entirely possible due to the changes in the strategic directions and restructuring of the business. For example, there are some airports that have changed their business model from that of servicing airlines to retail that requires a change to the value proposition.

In other cases, the value proposition might not change even though there are substantial changes to the business processes and the profit formula. This is often the gap that companies experience of which the management is often unaware. This is due to the changing nature of the senior management towards the use of tactics and strategies. Sometimes, instead of deploying tactics and strategy around the business, they change the business to adopt the strategy. Nothing is more obvious than the case of 'Here for Good' by Standard Chartered bank. The bank was a traditional British bank with strong branding in commercial lending and the savings account business. In the wake of the rebranding, the bank started to move towards brokerage and wealth management services that did not serve the bank's value proposition at all. Instead, the rebranding served to dilute the brand value by confusing the customer with its value proposition which, combined with a strategy that drives behavioural gaps, left the bank's reputation and balance sheet in tatters.

This is a significant area where further research will greatly assist in the application of Analytics in the context of business. For any young analyst, this is also the most important step in the application of BARCH and challenges the user of the model to discover more about the organization with which they are working. The Analyst needs to understand the value proposition, business processes and profit formula when they analyze a company.

Analytics

The definition of Analytics differs from dictionaries (Oxford, 2013) to practitioners (Sharma et al., 2010; Chiang, Goes and Stohr, 2012) but it is the systematic computational analysis of data or information resulting from the systematic analysis of data or statistics (Oxford, 2013). The nature of Analytics necessitates its positioning as an interdisciplinary area that integrates several business areas such as data management, database systems, data warehousing, data mining, optimization, and statistical analysis. The strategic value of business Analytics is the ability to generate insights and meaningful patterns in the data that has led to significant development in areas that analyse customer data (Kohavi et al., 2002; Davenport and Harris, 2007; Harris, Davenport and Morrison, 2010; Laursen and Thorlund, 2010). Because Analytics derives its root from science, statistics and

mathematics, the field is considered to be positivistic in nature. It is also strongly influenced by analytic philosophy that is also largely positivistic in nature and has provided the nomenclature of the field. To understand Analytics, we have to first understand positivism.

Positivism is a philosophical and research paradigm that originated from Germany in the 1920s. The paradigm was given a variety of names over the years including logical positivism, logical empiricism and neo-positivism. The key underlying assumption of positivism is the logical and scientific analysis of events or objects.

Positivism concerns itself with regular patterns and causal relationships existing in a sample of data. One result arising from positivism is causality. A causal relationship is defined as the scenario where two entities are connected by a relationship where the action of one causes an effect in the other. In research, this usually implies the if and only if scenario, however in practice, it is usually only an if statement. In practice, it is difficult to establish a strong causal relationship through data unless it is done through experimentations in rigorous manner. However, most research is observational leading them to adopt the weaker form of positivist approach which take the logical form of an if statement. Quintessentially, when a relationship is discovered such as X has an effect on Y, through Analytics, we cannot determine Y has an effect on X as well.

For Analytics practitioners, there are two sources of knowledge. These are logical reasoning and empirical observation. Any other sources are considered to be unreliable which is controversial among practitioners. This is because there is knowledge that cannot be observed directly or replicated scientifically but nevertheless plays an important role in decision-making and action. Secondary data in this case will also not be suitable for use, which is not the case for most research studies.

Logical reasoning is the derivation of insights and new knowledge through logical analysis such as formal logic. This type of knowledge does not require validation from empirical observations or experiments. The knowledge is absolute in nature and constructed from prior information that is also logical in nature. Knowledge derived this way is usually the result of formal mathematics.

Empirical observation works on the basis of researchers observing the occurrence of an event and deriving conclusions using these observations. Usually, most events could be analysed by empirical observation to come to some form of generalization, which is then supplemented by logical reasoning that could then be used to predict the next appearance.

One of the most important benefits of using positivism is the ability to formulate models that will lead to prediction. When the models are developed properly by thoroughly examining the events and data collected, and appropriate confidence limits are applied, the models can be very accurate. However, the approach makes it easier to predict that something will happen than to explain why it is possible to make the prediction. Positivism therefore assumes that there are independent causes leading to observed effects. Even though there is a direct relationship between the causes and the effects, positivism recognises this relationship but does not explain it. Explanation of an event, as opposed to the prediction of it, requires the use of deduction that can be difficult to apply. In Analytics, it is often less easy to observe and measure the direct variables that define a causality.

Positivists also attempt to operationalize variables where it is difficult to measure exactly what is going on. The positivist might allow a variable that can be measured to replace or represent a variable that cannot. For most Analytics practitioners, this is essential as the event that needs to be modelled is not easily measurable and proxies to the measures have to be created.

Another assumption common to positivism is the notion that the practitioner does not generate any effects on the sample population, and the actions resulting from Analytics models will not influence the outcome. The problem arises where the model is used on people. When a model is applied to individuals or groups of people there will be an immediate effect on the individuals or groups concerned. This reactance is a major issue in the application of Analytics.

Positivism therefore assumes there is an underlying reality, and encounters problems where there is no such underlying reality. For example, a researcher may be convinced there is a causal link between strategic alignment and the likelihood of medium-term company success. The positivist researcher assumes this reality exists and sets about finding it and proving it objectively. In fact, there may be no such underlying objective reality.

Validity is another major issue where a positivist approach is used in a social science or management application. Validity is the extent to which a piece of research actually measures what it is intended to measure. Empirical observations may apparently show this causality to be the case. The researcher establishes one causal link based on what he or she actually sees when, in fact, the causality lies elsewhere although it produces the same observations. Another key element that is important in positivistic approach is the need for reliability. The concept is concerned with the reproducibility of the results when a model is applied under similar circumstances. If this condition is met, the research is usually considered to be in a stronger position of causality. However, in the case of Analytics, the context of the problem makes it difficult for anyone to measure reliability. While it is highly desired, it is important to note that this is not always possible in every single case.

The positivist model is always needed in Analytics as the core principle is to develop a model which can be used for prediction and forecasting purposes. At the same time, models are required to establish a certain level of confidence in the causal relationship between the variables and the target. The mathematics involved also provide a sense of assurance to the users about the robustness of the model. However, given the choices of model, the analysts can almost always fit one of the models to the problem well and solve it. Even though the model may fit, it does not necessarily mean that it is the best model or the right model.

The concept of Analytics is to examine whether we have correctly applied the right model to the problem. Now it is almost impossible for anyone to be certain that the model that he or she has chosen is definitely correct when it is applied to a real world problem which does not have a solution. To solve this problem, we examine whether the model applied is powerful, appropriate and sufficiently robust. We will also need to evaluate whether we have exhausted all other modeling options.

Appropriateness in this case will be defined as the goodness of fit to the problem that needs to be solved. Now this leads to a situation where an overly complex model might be fitted to a simple model which fits wonderfully well with zero errors only for that model to fail spectacularly when

applied to other situations. At the same time, we have to be very careful building overly simplistic models which do not explain the complexities of the reality sufficiently.

Even if the model applied is appropriate, it does not imply that the model is robust. Robustness in this case refers to the model's ability to perform under various circumstances. In fact, this is actually something that is a major concern to the practitioner. A model might be robust in the context of the data that it was trained on and validated on. However, when it is applied elsewhere, the model may no longer be robust and changes have to be made to the model. This is usually achieved by re-training the model and re-validating the model again.

The relation between the power of the model with appropriateness and robustness is a negative one. The reason for this is that any powerful model relies on its ability to capture information that allows more accurate prediction of the target. However, this may imply that the model is using something that is too specific to the data that cannot be generalizable. Thus one can always build an extremely powerful model that works well with the data given but cannot be used anywhere else. To ensure the appropriateness of the model and its robustness, the power of the model will need to be weakened. In the case of Statistics and Mathematical modeling, there is a similar concept of Parsimony, which is commonly known as Occam's Razor. However, the two differs in the area of discussions. Parsimony refers to the need for simplest model with the best performance. However, such a model might not be relevant to businesses. Thus, the business' version is the tradeoff or balancing between appropriateness and robustness.

From the discussions about the nature of Analytics and positivism, I will define Analytics as the application of positivistic approaches to problems in a logical framework with the support of empirical evidence. Business Analytics is then the application of positivistic approaches to business problems in a logical framework with the support of empirical evidence. This conceptual view of Analytics derives from my prior experience dealing with the credit risk models in the bank. The bank wishes to minimize the amount of money, time and effort spent on building models. This is critical as constant redevelopment work can cause serious disruption to the application of the model on business. Models such as Credit Scorecards and Marketing Campaign models take months to develop and refine before they can be used. Once a model is deployed, it needs to be robust and able to deliver the value. At the same time, if the model fails, it will affect the campaigns, as they cannot proceed without correcting the model. The choice of model will need to be carefully evaluated for the success of the model.

One particular experience that I had in the bank about model building is the discussion of the concept of binning or discretization of continuous values. The practice is very common in the banks' Analytics teams and originated from older practices of model simplifications to make implementation of the model easier. However, with the improvement in computational resources and capabilities, this practice has been called into question. There were strong arguments from both supporters and detractors. However, the ultimate deciding factor that drives the retaining of the practice is the consideration of the impact on stability and robustness of the model. Through discretization, we can remove certain features in the data which occur as a result of some data quirks which may not exist in other time periods or data sets. This compensates the reduction in predictive power of the model for the stability and robustness of the model.

Another major challenge with balancing the aspects of the model is the difficulty convincing users of the benefit from the practice. A common problem is that the modeler can be over obsessed with 1 or 2 measures and all other aspects of the model are conveniently ignored. This again was experienced in the course of my tenure in the bank. There were extensive arguments that the model to be built was basically to achieve the maximum performance in a measure. For example, when it comes to the bad definition for credit products, some modelers advocate using the definition which produces the strongest model in terms of predictive power. However, this will cause problems as the definition might not make sense for the business.

The discussion about the application of Analytics is never ending but it is truly important for us to examine the assumptions that we use to build the model. A thorough examination of the model applied and the measures used will help us evaluate the likelihood of success using the model.

Revenue

Revenue refers to the gross sales or revenues of a company or organization and can be used as an allusion to any plans or actions that change revenue amount. In business, revenue is income a company receives through its normal business activities which is usually from the sale of goods and services to customers.

For non-profit organizations, annual revenue is also referred to as gross receipts. This revenue includes donations from the public and corporation, financial support from government agencies, revenue from organization's activities and revenue from fundraising activities and membership dues. In common usage, revenue is income received by an organization in the form of cash or cash equivalents. However, it is also critical to note that Revenue in the form of dollars and cents might not always be the sole factor in non-profit organizations. Such differences arise due to the nature of the organization which in some cases can only be measured by the impact of their policy on the wider population. For example, volunteer groups are typically measured by the impact of their services to community and not by the revenue that they generate. In these cases, they are considered as multiple revenue stream organizations where it is not just cash but also the overall policy impacts. The policy impacts are important factors for obtaining regular funding for voluntary organisations.

Revenue has also had to be pegged innovatively to non-profit objectives such as sustainability and social movement. This is commonly perceived in corporates as social responsibility. There are many interesting factors such as carbon footprint for logistics companies and philanthropic activities for technological companies. There are many corporates which have added these social responsibilities as revenue goals due to the positive effect they have on branding and sometimes the management of risk in ethically compromised scenarios as well as offsetting these activities against tax as in the UK.

No matter what business or organization is concerned, all require a steady stream of revenue to survive. Thus any analytical model applied has to be able to justify its existence and usefulness. This is because the model has to contribute to the company or organization and the most direct impact is

measured in the form of revenue. This approach also helps to make the return on investment easier and assist in the discussions with senior management. The use of revenue as a key measurement also makes the Analytics function appear as a revenue generating function rather than a cost center. In my experience, a lot of organizations make Analytics function as cost centers which reduces the appeal of the function. This is usually due to the stigma associated with a cost center which are minimized or even removed from an organizational function if possible.

Cost

Cost refers to the carrying value of goods sold. Costs are linked with specific goods using one of the many approaches that include specific identification, first-in first-out (FIFO), or average cost. Costs encompass all costs involved in bringing the inventories to place of sales or location of service. Costs of goods made by the business include material, labour, and any overhead. The costs of those goods not yet delivered will be deferred as costs of inventory until they are sold or written down in value.

Most businesses sell goods that they have bought or produced or provide services that they can provide. When the goods are bought or produced, such costs are considered to be part of inventory (or stock) of goods. These are treated as an expense in the period the business recognises income from sale of the goods. Businesses also have operations. These operational costs or expenses include those generated from rental of facilities, consumption of utilities and engagement of external business services such as accounting or web hosting. Most operational expenses cannot be avoided, as they are essential to the running of the business. In the case of service and volunteer or non-profit organizations, the cost will be primarily time cost or human hour cost that is additional to operational costs. Some non-profit organizations might also incur some material goods cost due to charity auction events, gifts or disaster relief supplies. It is clear that the cost of running a business of any kind is fairly complicated.

Determining costs requires proper records of goods or materials purchased and any discounts on such purchase. Any modification to the goods or service will require the business to determine the costs incurred in modifying the goods. Most modification costs include labour, supplies or additional material, supervision, quality control and use of equipment. While the principles for determining costs are easily stated, the application in practice is often difficult due to a variety of considerations in the allocation of costs and business needs. This problem is further compounded in the case of non-profit organization where the costs of the organization extend to items such as Man-hours or negative externalities of policies. Such cost are difficult to calculate and may not capture sufficient information.

Cost is always a major concern for any business and in the case of Analytics, this is one of the key concerns for new adopters of Analytics. From my experience, the reason for this is the need for costly hardware and powerful Analytics software. For Analytics practitioners, they need a reasonably powerful machine in order to process the huge volume of data required for their work. Such machines do not come cheap and every single practitioner will need one which further increases the cost. With the advent of technology, this problem has been alleviated by the use of cloud-based methods but it is not eradicated as there are some regulations on which kind of business

organizations are eligible to use clouds to work on their data. The other cost problem is the result of expensive Analytics software. These software are niche and cannot be replaced easily. Thus open source solutions are sought after by the organization as free replacements. Even though this has worked well in certain cases, it has also produced cases of catastrophic failure. An easy comparison would be with the pharmaceutical industry practices. In the pharmaceutical industry, it is important to have accountability as the experiments in clinical trials affect individuals and can be potential fatal. This result in a more conservative approach to handling the risk involved and they perceive the payout from death claims to be higher than that of software.

Human

This section considers the human aspect of Analytics that is mainly influenced by phenomenology and understandings of reality like social constructionism/constructivism. For any contextual based analysis, there is always the context which is often inseparable from the viewer's opinion and background. Since Analytics should deal with the human aspect of the problem, it is important to deal with Social and Human constructs. These constructs are often approached from several angles and disciplines. Depending on the construct, we have several possible ways to study, understand and solve the problem. Regardless of the nature of the problem, most of the approaches are essentially qualitative in nature and involve more exploratory studies. Phemomenology is most commonly used and is a sub part of qualitative research. These 'humanistic' approaches offeran alternative paradigm and perspective to positivism and consider the human relational dimension in how we construct reality and interact with context to be highly complementary to positivism. Humanistic approaches address several key weaknesses of positivism. They consider the complexities and incompatibilities of the real world. It is difficult to predict outcomes with fine accuracy where people are concerned due to the numerous possible human reactions to any given event. Every person has elements of unpredictability in his or her personalityand it is therefore difficult for positivismto assess or take full account of this uncertainty.Positivism can be unsuitable when applied to direct human issues such as motivation, individual and team development, motives and responses.

In my opinion the analytic practitioner must be like a phenomenologist who seeks to involve him or herself directly with the operations in order to understand the varied perspectives of stakeholders and grounds for support or resistance to change on any given aspect. Ideally the practitioner should become a member of the team that controls and operates the function. The more the researcher integrates into the team, the higher the level of understanding and appreciation of the operations that can better inform interpretation of data or prioritizing of data sets. The deep integration of the practitioner also reveals the constructs of the social fabric of the team that will not be reflected in data in general.

Phenomenology, as in social constructivism, also rejects the idea of there being a central underlying reality that exists and has to be discovered. The Analytics practitioner should see each event or activity as unique and as being driven by a one-off sequence or combination of drivers. In most cases, there is no explanation why the same set of drivers could not generate the same outcome each time.For example, the optimization of the nurse's ward scheduling is unique to that ward not

just because of the mathematical constraints but that it arises from the preferences and interactions among the nurses. A major aspect of phenomenology is that it is holistic. It emphases a much wider range of different variables than positivism, and it seeks to understand the complex linkages that exist between these variables.

Critics argue that the paradigm lacks the mathematical rigour and discipline of positivism, and it passively allows the researcher too much freedom of action. Positivist results tend to point to clear conclusions provided they are interpreted correctly. For the phenomenologist, the situation can be much more complex. The researcher has to interpret what he or she has experienced, and that interpretation is very much a function of the characteristics of the individual. The issue of replicability is very significant in phenomenology. It is very difficult to apply constants to aspects of human behaviour. There is a reasonable argument that phenomenology suffers from the issue of dilution.

Ethics is also an important aspect in the human dimension of any business.. Businesses always deal with the human factor either in the single structure or the group structure. This interaction requires ethical consideration at all levels of the stakeholdership. As business is an extension of the actions taken by the human elements, ethics needs to be core to the business as well. Revenue and Costs are generated by and through businesses and will be considered ethical if the business is conducted in an ethical manner. Ethics in Analytics depends on the human and business aspects, as they are the drivers for Analytics. However ethics is a major topic of discussion and deserves more than I can cover here and which I intend to explore in more detail as BARCH evolves. For the moment, the relevant issue for the purposes of BARCH, is the intended consequences and unintended consequences of the actions in BARCH. Given that the business is defined and can be examined for its ethical practices, we can ensure that the business is ethical for its intended purpose (Kenny, Pierce and Pye, 2012). However, we need to also consider the unintended consequences of a business that is set up for ethical work but manipulated for nefarious purposes (Slade and Prinsloo, 2013). An example will be bit coin, which was created as a secure payment service but manipulated and used by Internet black market traders of illicit drugs and weapons. Another serious example of Business Analytics practice raising concerns among practitising analysts and the public is the application of genetic material prediction models to identify the people who are more likely to get ill or are at risk from diseases (Juengst, 1995). However, such prediction models will naturally discriminate against people who have family histories even if they are healthy genetically. Even though the business of insurance is ethical, the application of such a model on humans will be unethical and discriminate against healthy individuals. The unintended consequences are severe for both the business and consumers. The business might lose reputation and potentially good customers. The consumers will be extended no insurance coverage or coverage at unacceptably high fees leading to wider social problems and medical care concerns. It is recommended that any user of BARCH has to constantly consider every single action both in terms of intended consequences and unintended consequences.

Given that the analyst is part of the human factor, it is essential that ethics forms part of the analyst's approach to Analytics. During the entire process of applying BARCH, the Analytics practitioner has to remember that the conversation and discussion needs to incorporate ethics. The direction that the Analytics practitioner drives in terms of conversation and discussion will shape the management as well as the organizational behaviour. Any ethical compromise will ultimately result

in decisions that are likely to be ethically unsound leading the organization and the Analytics applied astray. There have been cases where Analytics applied has led to ethically compromised situations that have caused much embarrassment[7] and in some circumstances, financial losses and damaged reputations[8].

One of the simplest tests commonly performed by risk specialists in the Business Analytics field is the front-page test. The front-page test assesses the impact of the model through the scenario of imagining it is made known to the public on a front page of media. If a person finds it disturbing to have his/her work on the front page of the newspaper or web portal, it is suggested that the project has most likely failed ethical standards of the public on whom the model will be applied.

Limitations

BARCH has its limitations as with any model. Being problem driven, the model does not attempt to develop any strategic plans. This could be a drawback for managers who are developing a business strategy though the use of business Analytics. While the results from the BARCH model can be used to develop a strategy, it does not develop the strategy directly. Even though the solution can be considered as a tactical solution to a problem, it does not address the overall strategic direction of organization. The model might generate insights to shape the strategy or guide the direction of formulation but is not the centerpiece of the strategy.

BARCH also assumes that any analyst user of the framework possesses in-depth knowledge of various aspects of the framework. The performance of the model is highly dependent on the skills and knowledge of the user. Being a framework, if the user does not have the right amount of knowledge, the model might not yield a solution. An example will be a prediction problem that requires a prediction model but is addressed by an optimization expert. The expert might not have the knowledge in the Analytics aspect to solve the problem and the framework will not be able to solve the problem.

The framework and the field require transdisciplinarity to work. If the Analytics practitioner is strongly biased towards one discipline, the framework might not yield solutions as the ideas and concepts from other areas are not integrated.

2.2 Combining Business and Analytics

Business Analytics is a new and developing field in computer science, statistics and information systems research. The strategic value of business Analytics has led to a significant development of business applications in areas that analyze customer data (Kohavi et al., 2002; Davenport and Harris, 2007; Harris, Davenport and Morrison, 2010; Laursen and Thorlund, 2010). Applications of Analytics in other areas like finance, marketing, production, manufacturing, human resources and research

[7] http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

[8] http://www.creditcards.com/credit-card-news/how-shopping-can-affect-credit-1282.php

and development have also been described extensively in the literature (Kohavi et al., 2002; Davenport and Harris, 2007; Harris, Davenport and Morrison, 2010; Laursen and Thorlund, 2010; Sharma et al., 2010; Franks, 2012; Stubbs, 2013). It has been argued extensively by advocates that organizations will generate a competitive advantage through the use of business Analytics (Davenport and Harris, 2007; Harris, Davenport and Morrison, 2010; Franks, 2012; Stubbs, 2013).

The concept of business Analytics is new in the information system as well as business studies literature. The transdisciplinary nature of the subject allows for various interpretations of the nature of the discipline depending on the definitions and problems. The various interpretations range from Interdisciplinary to Multidisciplinary to Trandisciplinary with various arguments to support or refute opposing viewpoints. The bulk of opinion is that the field is a Transdisciplinary field as the approaches always incorporate at least two (Business Analytics deals with Business problems and Analytics requires knowledge of a quantitative subject such as mathematics. Thus it will always have at least two disciplines) or more disciplines with results that will be insightful to both disciplines. In addition to the number of disciplines, the subject also requires a deep study of the context of the problem which is critical in the case of transdisciplinarity. Precisely due to this Transdisciplinary nature, it is hard for Analytics practitioner to identify the appropriate models or approach to apply and solve the business problem. The current frameworks do not address the business problem but seek to address the problem of managing Analytics. Within the literature, the most common opinion is that the new discipline is Transdisciplinary in nature.

The definition of Analytics in the Oxford dictionary is the systematic computational analysis of data or information resulting from the systematic analysis of data or statistics (Oxford, 2013). The Stern Business School defines Analytics as the study of data through statistical and operations analysis, the formation of predictive models, application of optimization techniques and the communication of these results to customers, business partners and colleague executives. Analytics can also be defined as the broad use of data and quantitative analysis for decision making within organizations which encompasses both query and reporting, and aspires to greater levels of mathematical sophistication (Davenport, 2010). INFORMS defines Analytics as the scientific process of transforming data into insight for making better decisions. There are also research papers which define Analytics as an interdisciplinary or multidisciplinary area that integrates data management, database systems, data warehousing, data mining, natural language processing, network analysis/social networking, optimization, and statistical analysis (Sharma et al., 2010; Chiang, Goes and Stohr, 2012).

Development in the field of Analytics has been driven by many experts (Davenport and Harris, 2007; Harris, Davenport and Morrison, 2010; Laursen and Thorlund, 2010; Franks, 2012; Stubbs, 2013) seeking to develop frameworks to assist senior management with strategies and approaches for the application of Analytics in their companies to realise a number of ideas and visions that will keep them at the leading edge of the market. However, the limitation of these frameworks so far is that they are generic in that they are useful as strategy guiding and planning tools but they cannot be applied to more specific cases that can be operationalized into an Analytics project. The frameworks

were originally developed in the earlier stages of the discipline when it was unclear how the discipline can be incorporated into business practices. With the evolution of the subject and as Analytics become more common among organizations, new users and analysts need to be able to identify areas where they can apply Analytics and link the results from Analytics back to the business in a beneficial way that helps the company and improve its profits (Franks, 2012; Stubbs, 2013). To achieve that requires the Analytics practitioner to have knowledge of the specific business processes without which a difference cannot be made quickly and efficiently through a pre-tailored approach.

In the next section, to contextualise my field of practice and to highlight the contributions my work has made, I review business Analytics models and discuss their strength and weaknesses. Practitioners in the field of Business Analytics have developed all the models that I review. While there are other frameworks such as CATWOE and CRISP from other fields adapted to the business Analytics context, these models are not transdisciplinary in nature and since they are merely adaptations, they do not capture many of the learnings from the practice. Contrasting my work with the work of other practising authors in the field of Business Analytics is a very important and rewarding way to extract more learning and to enhance my articulation of BARCH and the rationale behind its development.

3. Business Analytics Models

3.1 Business Analytics Model (Laursen and Thorlund)

The Business Analytics Model is an enterprise level Analytics Implementation model proposed by Laursen and Thorlund in the first chapter of their book (Laursen and Thorlund, 2010). The authors describe the model as an overview of a large and complex project with many people with a range of skills and competences with an eagle eye perspective. The mixture of individuals with such diverse skills necessitate the transdisciplinary approach as their combined capabilities drive the success of the business. The key objective of the model is to deliver an outline for understanding and developing successful Business Analytics in any environment. The model also provides a single point of reference for the overall structure contributing to a successful Business Analytics which also clarifies the roles of contributors, the creation of information, the flow of information and the consumption of the information.

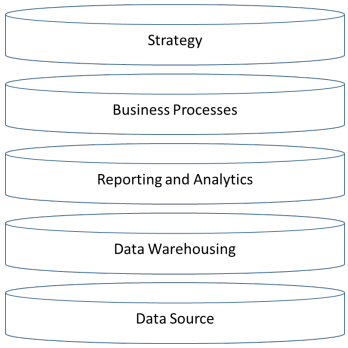


Fig 1. Business Analytics Model

The model has five layers. The first layer is strategy creation, which involves the selection of information strategy that will be relevant to the organization. This is also the top layer and likewise is determined by the top management, as they are the ones that will be making and designing the overall strategy for the company. However, this particular layer also presents a problem for Analytics practitioners using the model. The model seems to assume that the management already have an objective in mind when they want to implement an Analytics solution. At the same time, even if they design or select an information strategy, it does not mean that the strategy will be suitable for Analytics. The creation of strategy also assumes that a strategy exists for the company to use and there are situation where appropriate strategy might not exist.

The second layer is known as Business Processes and is directed to the operational decision makers who support the company's chosen strategy. The main role of the second layer is to determine the relevant information and knowledge that is needed to support the chosen strategy. Information and knowledge are not the same. Information is opinion or data in the organization while knowledge is information that has been tested and determined to be useful to organization. In classical terms, information is more akin to the concept of Doxa, which are primarily information that is available. However, this is not commonly accepted knowledge like Endoxa, which has some kind of validation to verify its authenticity. Gnosis is more akin to the concept of knowledge although it is noted to be more personal in nature and, in the case of organization, something more organic to the business in question.

The first two layers primarily involve the management and look at the needs of the organization. They focus on how Business Analytics will be able to support the business and Business Analytics required to achieve the successful implementation and achievement of the goals. However, business processes are often not in line with the company strategy and require alignment of processes. This process can be extended over long period of time. Even if the processes are determined, there might not be data to support the Analytics needed. The knowledge of the processes might be flawed and information is incorrect which can lead to problems. Determining useful information and knowledge by itself might require extensive application of Analytics as well. Given such complexities, I do not consider this layer to be independent of other layers. BARCH approaches the strategy and business processes differently. The Business Analytics model focuses on the strategy of the company which ultimately linked back to the business nature and its goals. If the business goals do not fit the business' nature, then it will have a negative impact on the business. To prevent that, BARCH emphasises on the importance of business in driving the two pillars of revenue and cost taking into consideration human factors and Analytics model.

The third layer is also known as the reporting and Analytics layer. This is the bridge layer or transformational layer that links the first two layers with the next two layers. Being right in the middle of the five layers, this is also the layer that is managed by the middle management as well as the base layer of analysts supporting the Analytics. One of the key roles of the layer is to generate analyses and reports using information fed in by the next two layers. This is where information gets generated and validated as knowledge. However, as mentioned in the previous layers, this layer is subject to the presence of the previous two layers being placed in a suitable combination with this layer to generate value and solve business problems.

The fourth layer is the data warehouse layer which is involved primarily in data gathering and processing. The key focus of the layer is to ensure that all the data are held in a single place and processed in a way that is easily accessible to users such as analysts. At the same time, the data warehouse must hold the data in a way that any users will be able to understand the data without having to infer on their own. The data should be cleaned and standardized to make it easier for users. While in the original model, the fourth layer is placed beneath the third layer, the layer is actually laterally linked to the third layer. This layer is common among organizations that have existed for some time. In the case of start ups, this is unlikely to be present and this layer might need to be the first layer for certain organizations. This is the case for the start up that I dealt with where they did not have any data to begin with. However, this does not mean they cannot be using Analytics for their business. They can use Analytics techniques that do not require huge amounts of data to work with. This is one area where BARCH is superior in the sense that we use Analytics as something to determine the combination of data and technique instead of data dictating the techniques.

The final layer is the data source layer which is concerned with data source and IT infrastructure. The layer's key function is to ensure that data are collected from the daily activities of the business and the IT infrastructure is able to handle the data collected and support all the Analytics requirements of the organization. The IT infrastructure does not just handle hardware alone, it also handles software. In the original model, this layer is placed below the data warehouse layer. However, the layer is more laterally linked to the third and fourth layer. This is almost a universal layer that every framework will need to consider in one form or another.

3.2 PADIE

The PADIE framework or technique was first described in a white paper by IBM (LaValle, 2010) which addresses the development of Analytics and its implementations. The framework was developed specifically to address the difficulty for business to understand Analytics and how the implementation can be delayed by this. The authors proposed the use of the Process-Application-Data-Insight-Embed (PADIE) technique which will enable the business to rapidly deploy Analytics by embedding it into the business process.

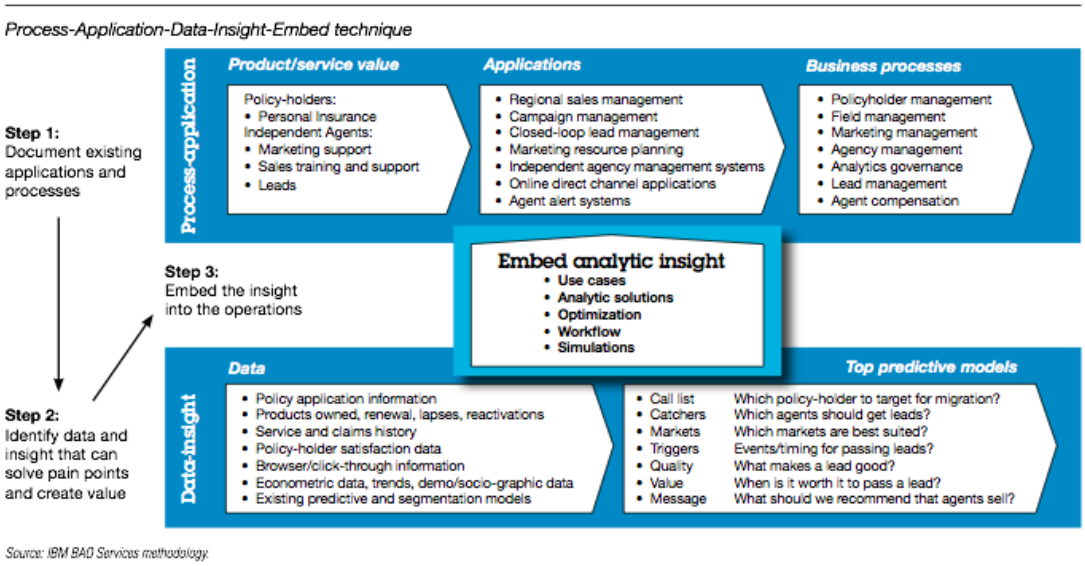


Fig 2. PADIE Framework

The PADIE technique consists of three steps:

1. Documentation of existing processes and applications
2. Identify data and insight that can solve pain points and create value
3. Embed analytic insight

In the first step, organizations identify the value they deliver to customers, the applications that drive the business and core processes such as management systems and metrics. Typically, in this step, there is a detailed review of the business processes and how these processes are linked to the business drivers. The metrics and key performance indicators have to be adjusted to link the drivers to performance. In this particular step, the key objective is a thorough understanding of the business.

Once the first step has been achieved and the business processes are transparent and linked to the business drivers, the organization begins the second stage by identifying the questions – who, what, where, when, why and how – that will address business issues and create revenue, cost or margin value. The key objective here is to channel the efforts of the modelers to drive their analytic inquiries into the data which will result in impactful results for the organization. Organizations also need to identify key data sources that will be used during the analysis.

The last step is the most important step for value creation. The organization needs to determine the best approach to embed analytic insights into its operations. Organizations have many choices regarding the embedding approach which includes using cases on application enhancement, new analytic models can be introduced, optimization algorithms combined with existing rules engines, new business processes and using simulation studies to enhance management's ability to understand various scenarios. These successes through embedding insight into processes can make a lot of difference to the success or failure of Analytics initiative.

The PADIE framework is very loosely related to Analytics and merely overlays it over business. It does not attempt to link the business with the results and Analytics. The approach is also very process driven and over emphasises the need to embed the Analytics process into the business. While this is good, some businesses can be very fluid and the process of embedding Analytics can be disruptive. Very often, business process can change due to business environment or internal constraints. If the business Analytics models are directly embedded in the process, then it will require remodelling whenever there is a change which can be a severe drain on resources. In the BARCH model, the model is merely a supporting tools and not the guiding model behind the business as it depends on the business and not the other way round. There is also the danger that the company's processes are not well documented and thus create delays due to the need for proper documentation of processes. The PADIE framework also does not recommend the kind of measures needed to establish the success or failure of Analytics. BARCH on the other hand prioritises the business and not the existing processes that can be flawed or non-existent.

3.3 DELTA Framework (Harris, Davenport and Morrison)

The DELTA framework is one of the most popular frameworks in terms of citations in journals and Analytics practitioner publications and references by Analytics practitioners and was created by Harris, Davenport and Morrison (2010). Its popularity is achieved through the thoroughness of the model for strategic thinking as well as the applicability to management requirements for implementations of Analytics. The framework is also often cited for defining the values that Analytics contributes to businesses (Kohli, Rajiv, and Varun, 2008; Ko, Lee and Lee, 2009; Shuen, 2008; Rousseau, Manning and Denyer, 2008; Chen, Chiang and Storey, 2012; Nudurupati et al., 2011).The framework assesses the capabilities of the company in terms of its capabilities to achieve greater success through the use of Analytics. The model has been discussed mostly in the view of the senior management (Shuen, 2008).

The five letters represent the following aspects:

- D for accessible, high-quality data
- E for an enterprise orientation
- L for analytical leadership
- T for strategic targets
- A for analysts

Data

Data is placed first as it is the prerequisite for everything related to Analytics. For data to be useful, it has to be "clean" in the sense of reliability and validity. The most common example cited in literature (Davenport, Harris and Morrison, 2010; Franks, 2012) is the Customer data. The customer

data has a unique identifier for each customer with the relevant customer names, addresses, and purchase histories that are generally reliable and valid. Even though the field meanings and uses are commonly understood by users, they require proper validation and thorough understanding. When drawn from several sources, the data has to be integrated and consistent for it to be of value to the users or analysts. To encourage use of the data, it has to be accessible in form of data warehouses.

To understand the ideal scenario for data that are employed at the most analytically sophisticated companies, the literature describes the requirements and suggests that while the ideal case will facilitate most Analytics, there are several key components of data management for Analytics that are common to the most sophisticated companies. The authors of the model suggest that it is not essential for every organization to be at stage 5 for data. This critical aspect of data management is managed through a careful evaluation of how close the organization can come to this ideal.

The key steps to the ideal data management scenario is as follow:

- Structure - the nature of the data
- Uniqueness - exploitation of the data
- Integration - consolidation of data from various sources which may or may not be compatible
- Quality - reliability of data
- Access - obtaining the data
- Privacy - protecting the data
- Governance - controlling the processes

Most organizations, as described in literature (Davenport, Harris and Morrison), store their data in one of these forms: Cubes, Arrays and Non-Numeric. The nature of data storage will affect the types of analyses you can do. Most transactional systems store data in the form of tables. Tabular storage excels in processing transactions and creating lists but is considered to be less useful for analysis. The literature describes this weakness as a result of storage limitation which spans 12 months in most scenarios. Data is often extracted from the database or transaction system and are stored in a warehouse mainly in the form of cubes. Data cubes are collections of pre-packaged multi-dimensional tables. These data cubes can have many dimensions more than three of which can make them difficult for most people to comprehend. However, they are useful for reporting and data exploration work such as "slicing and dicing". This is facilitated by the limited number of variables used which are important to the analysts and the final report.

Data arrays are also tabular form of structured content. In this format, a particular field or variable can be used for analysis if it is in the database. Arrays may consist of hundreds or even thousands of variables. This format maximizes the flexibility of the data for analysis and reporting work. However the format may be confusing to non-technical users who do not have the same understanding of the database especially with regards to structure of the database or the source of the fields of the data.

Unstructured and non-numeric data which is the current frontier of Analytics, are not in the formats or content types that databases normally contain. This type of data can take on a variety of forms and companies are collecting massive amount of such data which has increased their interest in analyzing it. The phenomenon is occasionally called the Big Data Phenomenon. This common association is due to the velocity in which data is collected which has directly caused data volume to

increase. Due to the unstructured nature of such data, the data comes in many forms and this result in variety. The data may be from external sources such as web pages, blogs and twitter where each differs from another due to the way the data are captured. In the case of internal data, these are usually textual data from various systems which can be items such as warranty or customer support emails. These various data sources require proper storage in order for them to be useful. Very often, the data required are textual data and specialized storage has been developed for such data (Hadoop, 2010; Hive, 2010).

The "potential value in unstructured data is incredible" has been described by influential publications (Chen, Chiang and Storey, 2012). However, mining for data in unstructured data is similar to mining for gold where the insight is embedded in plenty of noise like gold in sludge. An example will be statements like "Have you seen the Sun in Britain?". Such a statement is ambiguous due to the nature of the target which can refer to the newspaper in Britain or the Sun that we know of. Given these confusions in the textual context, it is crucial to execute semantic analysis to extract the proper meaning. Most stage 5 organizations are well engaged in numerous projects involving both cubes and arrays. However, they have only begun to use or experiment with unstructured data like images, Web text and voice analyses.

To exploit data and outmatch the competition, one needs to have access to data that is unique to the company or industry. Given that companies with similar Analytics capabilities are able to tap into most data, should the data be the same, it is likely that similar conclusions will be arrived at. Achieving an analytical edge will require some form of unique data. The importance of measuring the value of the information and collection of proprietary data that have not existed in the organization cannot be underestimated and requires extensive planning. With unique data, it is possible to develop a unique strategy. However, even though it is possible to gain the initial advantage, the competition may eventually catch up and that is when innovative data collection will help in maintaining that edge (Chen, Chiang and Storey, 2012).

In the entire discussion about the types and nature of data, the DELTA framework is fairly specific about the importance of data and how it can be used to deliver great value to an organization. The authors also specifically dealt with the need to collect all sorts of data and that data is the first step to an enterprise wide Analytics implementation. However, data is the product of the business. The framework does not recognize that while data is the key to building Analytics practices, it does not address Analytics problems. The discussion about exploiting data and leap-frogging competition is all about identifying problems or areas of potential which requires a close examination of the business. Data is useful to point to the areas if the right data is collected, but in the case where data is not collected, it is still possible to identify the areas generating problems. This is a problem of Chicken and Egg whichever comes first. In the case of Analytics, business must always first exist before data. Unlike DELTA, BARCH emphasises the importance of business as the driver of Analytics. In a way, the BARCH framework has a more "Meta-Business" view than the DELTA framework.

Integration of data from multiple internal and external resources is a very important aspect for analytical organizations. Most of the transactional systems in organizations do not allow flow of information between functional areas and they are mainly focused on particular areas of business. The advent of Enterprise resource planning systems enable the integration of data sources and helps to alleviate the data integration problems in legacy systems. While ERP systems alleviate a fair bit of

data integration problems, they do not remove the need for proper data integration. Many organizations will need to consolidate and integrate data from web or external data sources and providers with the internal system.

Sometimes, there is a need to integrate third party information with internal information in order to handle customer orders. There will be a strong need for proper data integration. This scenario is most commonly encountered by retailers and banks. For banks, credit bureau access is frequently needed for credit operations. Unless the social security numbers are given, data integration can be very difficult. In the case of the retailer, it is important to reconcile online or web orders with internal inventory data in order to ensure that the orders can be delivered on time. Given that data silos are found in many organizations, data integration is unavoidable and a key asset to companies who are at Stage 5. Through data integration, these companies have access to different data elements such as customer information, product inventory, supplier quality and others.

Unfortunately, data integration has been widely touted as an unglamorous task that has been criticised by many. Very often, the task involves conversations with several parties and this leads to discussions about data quality. Such discussions often lead to unpleasant squabbles between the parties. At the same time, data integration often comes under Master Data Management that can be overly complex and ambitious resulting in lack of results and poor final products. While it is important to have good data, it remains a fantasy for many large organizations.

Even though data quality is important in analytical decision-making, it does not need to be perfect compared to transactional systems or reporting applications. There are techniques available for analysts to deal with missing data even in the case where there are substantial amounts of missing data to get around the problem. Nevertheless, flawed or misleading data remains a major problem for Analytics especially if the problem is reliability as opposed to validity. Unlike reliability, validity can be resolved in the data through proper data validation checks. Reliability on the other hand has no such checks available. This is why having integrated data is important even if each data source has its inherent problems.

Armed with a business objective and the choice of analytical model in mind, analysts will need to trace all the data problems back to the source that is usually to the data entry point where the data was originally entered. This allows the analyst to find the root source of data error, discern the solution to the problem and to fix incorrect data. Even with the best and most sophisticated ERP system, it cannot prevent front-line employees from entering incorrect data.

Analytical companies do not always have perfectly clean data, but they address the data quality problems to ensure sufficient data quality for Analytics that affect business. Most analytical driven companies have a well-defined and relatively painless process for improving quality of data as needed. These processes ensure proper capture and validation of data to reduce cleanup work.

Before analysis can be done, access to the data must be given. Most Analytics driven companies control the access to data by using the data warehouse or data mart. These data warehouse or data mart are usually implemented on an enterprise wide level. Such data warehouses or data marts are all encompassing and flexible enough to incorporate external data sources. In the case of big enterprises, the data volume may be so big that they overwhelm the people working with them. To

alleviate the issue, most enterprises will have smaller data marts to cater to individual needs. These data marts are functional driven versions of data warehouse occasionally created independent of IT.

Most organizations with strong analytical approaches tend to have information about the entities that are important to them. These organizations have well thought out and thorough policies with regards to the privacy of information. Given the cross-national boundaries of many such organizations, they often have to adopt different policies in order to meet the country requirements such as the Data protection act in the UK. They protect their data from unauthorized access and data held are often done with the explicit permission of the owner of the data. Usually, they are conservative with their data control policies.

The purpose of governance of data is to ensure the organization has a structure in place to ensure consistency, reliability, standardization and access. Even though it is the responsibility of the organization, proper assignment of duties will be essential to ensure that ownership of the data is shared among key stakeholders. Very often, the senior management is the key driver for alignment of data to the organizational goals. The senior management has to define the various drivers for business in terms of data and collaborate with one another. Most management teams do not discuss information in this manner, but without such in-depth discussions, the enterprise wide data integration will be difficult. High-level decisions about data have to be made by senior management especially when it deals with issues like ownership, stewardship and most importantly the relationship between data and strategy. The relationship between data and strategy will ultimately decide the fate of the organization.

The premise of data being the pre-requisite of everything in Analytics is unfortunately a false premise. The practice of Analytics is the method of logical analysis and thus independent of data. Everything about a problem is merely the lack of a coherent model or framework to explain the phenomenon that is backed by data. This lack of suitable data should lead to greater search or collection of data or information that is then reflective of problem and can be used to solve the problem. However, the DELTA framework positions itself as something that requires data before the subsequent components can function. This places data as more important than the business that it reflects. BARCH places emphasis on the business for which we are trying to build an Analytics model. This approach focuses on the effort to understand the complexity of the business that ultimately gives meaning to the data that it generates. Focusing on data is a little myopic in that sense.

The DELTA framework also constructs several criteria for the ideal scenario for data. However, I found the framework lacking in practical usage in real life situations. For example, discussion about the nature of the data quickly descends into sources structure and how tabular works differently from other forms such as unstructured data. This is futile, as data may exist in too many forms to be practical for anyone to truly categorize them properly for work. However, we need to understand how the data relates to business and how they came to be in this form.

Uniqueness also poses an unusual demand on the data. Since the data generated is fairly unique to that business, it is hard to assume that the data will be different. In the DELTA framework, the discussion is focussed on the collection of information that is not commonly available to competitors who in my opinion will be the case for most organizations. Of course, one could argue that there are other more unique data to be collected. I think this ultimately ties back to the business. If one is not in that area of business, how will irrelevant unique data be useful? Thus, it is important to

understand the uniqueness of the business and how we can exploit data that complements the business. Issues such as integration, quality and access will naturally fall in place once we have a good understanding of the business and how we can exploit those data.

Governance is an issue that is very important to Analytics which is emphasized by DELTA framework. However, it is ultimately a business issue rather than an independent data issue. Most businesses practise data protection. However, in the discussion on data, governance is given a strong prominent position. There are plenty of discussions about the relationship between data and strategy among Analytics practitioners particularly on the issues of whether Analytics should be managing compliance and control issues that are non-mathematical oriented subjects. The nature of Analytics and debates further strengthened the position that discussion about governance of data on an enterprise level is inappropriate for Analytics that is better served by a debate on strategy and Analytics. The management especially during meetings with the senior management is often focused on defining key performance indicators (KPIs) and how they link with strategies and data backing those strategies. This discussion can also be found in the subsequent section on Targets. In BARCH, data is the result from business processes that will affect the types of Analytics that can be used and is related to measures of Revenue and Cost. The BARCH model does not see data as an independent item but rather the product of business and strategy. The data can also be used to verify the strategies' efficacy. The relations of data with business, revenue and cost are clear and reflect the business as a whole.

Enterprise

The literature discusses the challenges of data management (Davenport, Harris and Morrison, 2010; Franks, 2012) and on the whole suggests they are answered easily if the enterprise controls the important data, analytical software and talent. More importantly, the management must be motivated to cooperate on analytical initiatives starting from data. The conditions are important for three reasons:

1. Major Analytics applications that deliver improvements to performance and competitiveness will concern multiple sections of the enterprise due to the need to incorporate the various aspects of business.

2. If applications are cross-functional, it doesn't make sense to manage key resources—data, analysts, and technology locally. Given that data can be collected from multiple sections of the enterprise, the application of Analytics is therefore multi-faceted and requires resources from those sections. This is often neglected as the focus for most projects are targeted at certain business functions and the initiatives are considered in silos.

3. Any Analytics project without a scope that takes into consideration the enterprise perspective will unlikely deliver results that have an enterprise reach.

To develop an enterprise view of Analytics, data integration, pooling of resources and the development of a corporate IT platform is the base level for any organization to succeed. To advance to the enterprise level Analytics, any fragmented efforts and constructs of individual managers will

need to give way to a single, holistic view of the company. This creates an effective management practice that can then address the strategic issues at the core of business performance and organizational competitiveness. Such issues include:

- What impact do the performance factors have on future growth and profitability?
- How can we forecast volatile market conditions and control the fluctuations?
- How does the allocation of investments across product geographies, and marketing channels optimize the return?
- Are management's decisions in good alignment with a company's vision and objective(s) or are they aligned with self-interest?

Using Analytics, it allows Analytics to facilitate management's ability to understand the business on an enterprise wide level. The enterprise level coordination also ensures that there are good and strong standards with regards to the practice of Analytics. The enterprise level approach also allows the development of an Analytics roadmap that provides directions, requirements and values of the projects that will then add value to the organization. This road map also prevents people from setting up fiefdoms where there will be duplicated efforts and struggles among the analysts. An enterprise wide view will result in a roadmap where the IT efforts are channelled and the complexities of implementing Analytics will be reduced.

The Enterprise aspect of the DELTA framework is again strongly related to the business aspect. This aspect of the DELTA framework focuses on the development of the Analytics capabilities on the Enterprise level. The approach is to have the senior management organize the efforts and direct the integration of information to form an enterprise view. This is related to the data aspect of the DELTA framework. However, this aspect does not explain how it directly contributes to the application of Analytics and effort to improve the business. The roadmap discussed is more akin to an organizational restructuring than addressing concerns of a business. Unlike BARCH, this is not a relevant activity to any business. The assumption that senior management backing is present can be easily challenged and this will render this factor nullified. From another perspective, even if the DELTA framework purely serves as a form of reference to development of the capabilities of an organization in Analytics, it needs to take into account the situation where senior management is not on board for that.

However, the BARCH model, views the business as the key focus area. This approach ignores the requirement of engaging the senior management and concerns the group of Analytics practitioners applying the model. This makes the application of Analytics easier. Without the discussion about the roadmap and restructuring, this makes the introduction of Analytics less intrusive for the organization. The linking up with performance factor is natural for BARCH as both revenue and cost are crucial to any business and definitely important to the C-level executives whose KPIs usually involves these measures. This helps the adopter of BARCH to link better with the management.

Leadership

In the DELTA framework, Leadership is the next major component. It is important that Analytics gets invited into the organisation in the first place and can be achieved only through the decision making process of senior management. In order for organizations to fully capitalize and integrate Analytics in their decisions and business processes, there is a need for special leadership. According to Davenport and Harris (2010) and Frank (2012), such senior management are not just convinced by analytical projects they have seen being successful. They harbour great passion for managing business through the use of fact based decision-making process.

Leadership is the single most important factor that will determine how analytical an organization will be. Their support for analytical initiatives will most likely bear fruit and they can strongly influence the culture and gather resources to move towards Analytics driven decision-making. Even though the CEO remains an important figure for a full-fledged Analytics competitor, leaders at every single strata of the organization have a role to play and move the organization towards the goal of Analytics competitor.

The key attributes of the 'special leader' can be common across organizational levels. Very often these traits are crucial in the development of the skills and expertise among the users of Analytics and promote the use of Analytics.

- Development of People (or Soft) Skills
- Motivated to Gather more Data and do more Analysis
- Hire the relevant talents and acknowledge their work
- Lead by example, lead them into battle.
- Result Driven
- Educate and Develop
- Develop Strategy and Performance Measurements
- Go for Domino Effect
- Prepare with a Siege Mentality
- Build an Analytical Ecosystem
- Be prepared for Multi-Pronged situation

The DELTA model has an extensive discussion about leadership and how it is important to have the necessary skills and personality to deliver the values of Analytics. However, not everyone is a born leader and most people are not leaders. To make Analytics accessible to businesses, we need analysts and not just leaders. While leaders are powerful in driving the Analytics in organization, it is always easier to convince the management with cold hard numbers and return on investment than a charismatic leader who fits the profile.

The role of leadership is important in every aspect of a business and not just Analytics. For most implementations of process and policy in organizations, senior management and key leadership figures need to put their support behind Analytics in order to make things happen. While arguable that the leadership is critical in DELTA, it would be disastrous to think that the leadership alone can drive the entire organization to Analytics. Even Alexander the Great had to concede and retreat when his men refused to fight anymore[9]. Any project implementation should instead focus on the humans involved in the project as they are always the stakeholders and the driving force behind the

[9]http://en.wikipedia.org/wiki/Alexander_the_Great#Revolt_of_the_army

project. BARCH focuses on the human aspect that is more general and contemplates the relations between the stakeholders. In this aspect, BARCH presents a more balanced picture of the dynamics among the users and stakeholder so that we can work on any possible conflicts.

The leader who has already started testing Analytics would have developed some of the skills mentioned and naturally planned for an organizational level deployment of Analytics. Some of the traits may not be as relevant to actual analytical works. Other traits are mainly ideals of leadership generally and are not restricted to leadership in Analytics. The discussions on leadership is unfortunately inappropriate in the discussion of Analytics and should be discussed in more depth in management books especially where Analytics is becoming increasingly used to increase the competitive edge in global business. Analytics is about problem solving and not management practices. While it requires an analysis of the business which will involve leadership, the key is not to focus on the traits of great leaders which is in the domain of management practices.

Targets

Businesses need targets to drive their strategy and allow them to justify the use of Analytics through results. The target may be any aspect of the organization for which Analytics may be deployed. Targets are needed because there are not enough resources for all aspects of the business and it is not easy to be analytical about all aspects of businesses.

It is important for an Analytics driven company to focus its analytical efforts to achieve the most good given constraints in resources. Given the varied business opportunities, only selected ones will confer the breakthroughs in performance or differentiation in the marketplace that is highly sought after. To begin using Analytics, addressing a specific and critical business problem will be a good initial target. Once experience has been accumulated and rewards are realized, the targets can cover broader areas which allows for more strategy to optimize key business processes and bring innovation to the operations that differentiate the business in the eyes and experience of its customers. A good target is very important to the business especially if it engages top management, solicits their commitments and creates opportunities.

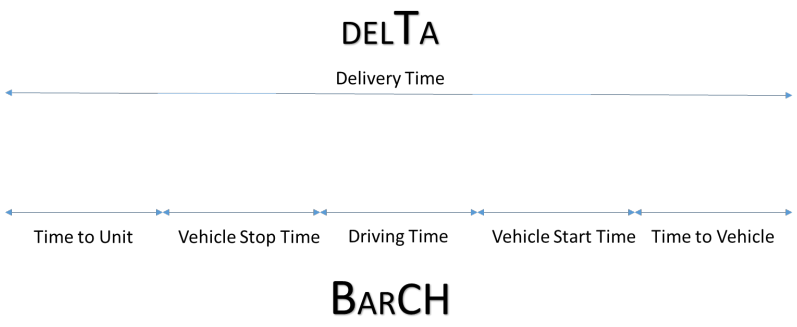
The organization's overall strategic plan is all about finding opportunities in areas such as business growth, innovation and marketplace differentiation. Until the strategy incorporates the use of Analytics and details the targets, it is important to discover analytical opportunities in the business. A simple approach would be to consider the new ideas and practices used elsewhere in other industries or companies in the same industry. It is also important to evaluate the need or opportunity to evolve. We must also be alert to the activities of competitors on a regular basis and be abreast of the developments in the industry analytical applications. It is important to innovate to differentiate from the competition.

It is important to understand how business performance can improve and what the key factors that drive performance are. This requires exploration about the business; understand the fundamentals and where will the next breakthrough appear. The most common approach is to conduct a

systematic inventory review of key business processes, the decision making process within them, and the business decisions which could benefit from more and better Analytics.

In this aspect, BARCH and DELTA share the same concerns about setting targets but they differ in some ways. The discussion of Targets in DELTA is fairly generic and does, at times, mention the strategic direction of the organization. BARCH is more tactical and focuses on the key KPIs of Revenue and Cost that are critical to every organization (Choy et. Al., 2012). In my work on the Key Performance Measures of Parcel Delivery company, I highlighted the importance of several key KPIs common to the business that are appreciated by the management. However, the management were quick to recognize that I raise more measures than were expected.

The longer than expected list of measures is the result from the use of BARCH. When DELTA is used, it is done at a strategic level which can be driven primarily by cost or revenue factor. Due to this strategic level drive, the management were even thinking of 5 minutes intervals between deliveries without considering whether the targets looks ridiculous when they hadapplied the DELTA framework. By tying them back to measures involving business, the familiarity of the measures formulated from BARCH helps them to better review the targets. The focus on the business also forces the management to review measures which cannot be possible with the business. An example is the interval between deliveries.



The entire approach using DELTA is plainly insufficient in this case. The DELTA framework will take the measure of delivery time and attach a value to it in relation to the business. The construct does not appreciate the detailed information of operations which results in a measure that is detached from reality or business. When BARCH is applied, the aspect of business is perhaps the most crucial one as we need to appreciate the process of delivery as it is a complex procedure. Through the business process, we break it down into smaller components which are similar to costs and how humans view the process. This relates the business to cost and human components which increases the reality of the measure and its relevance. The example also demonstrates why most discussions of Targets in the DELTA framework are too broad. BARCH focuses only on tangibles which can be measured and how to convert intangibles into measurable forms.

Analysts

Analysts have two chief functions. They are the creator of models that facilitate the achievement of targets and they also bring Analytics to a level that enables business people to appreciate and apply them. Analysts are the human aspects of business. The key aspects of analyst lie in the skill set of the analyst, the importance of proper management to ensure motivation and the organization of the analysts to allow them to perform to expectation.

The DELTA framework discusses the analyst in terms of their core skills and how they bridge the management, business and Analytics. The framework discusses the needs of the analysts and how to motivate them in a way that is consistent with the organization's goals. This is critical as humans are an essential part of any organization. The human aspect is also part of the BARCH framework. However, the frameworks differ in the interpretation of the human role in Analytics. DELTA engages in it by looking at strategy and organizational behaviours. BARCH examines the relationship of human factors to the success of Analytics. There are overlaps in the topics of discussion but they refer to different areas. However, in my opinion, the emphasis on strategy and motivation in DELTA subordinates its importance to a sub development of business processes. This is in complete contrast to BARCH which positions human interactions as the key element to determining the success of Analytics.

3.4 G.R.E.A.T model

Bill Frank (2012) in his seminal book on taming the big data wave, came up with the concept of the G.R.E.A.T model. The model is created and formulated on the basis that for analysis to have an impact, it has to be conducted in a particular manner. By doing it according to the model, the various factors will interact together to make the analysis great. The core principles in this model are Guided, Relevant, Explainable, Actionable and Timely.

Guided

Any great analysis has to be guided by business needs. If the analysis is not guided by the business needs, then the analysis will not be useful to business and it is done just for the sake of it. With access to huge amounts of data, it is very easy to be side tracked to attack interesting problems rather than those that are crucial to the business. To produce a great analysis, one must begin by identifying the specific business problem that needs to be solved. Once identified, the analysis can then be guided by its relevance to the problem. This way, the analysis will be tuned to the overall objective that is specific to the business problem being addressed.

Relevant

Since the analysis is guided by business needs, it has to be relevant. But this means more than just randomly selecting some business problem. The problem has to be one acknowledged by business to need an urgent solution. At the same time, the problem has to be one that can be addressed by the business. At the same time, we have to make sure that the result will be relevant for to the business for a time period if not forever. An analysis that is only relevant for the next 5 minutes will not be relevant for a business decision 3 months down the road.

Explainable

Even if the analysis is guided by business needs and is relevant, the analysis must be easily comprehensible to the business users. While modellers are mathematically oriented and comfortable expressing themselves with algorithms, pseudo-codes, mathematical proofs and statistical formulas, such verbose use of complicated symbolic logic serves to obfuscate the simple nature of the analysis to address the business problem. All the technical details are needed for proper auditing of the work, but with results, they have to be simplified in ways that can be expressed by non-technical trained folks to the business users. Richard Feynman once said that if one truly understands the model, they could explain it to a kid.

Actionable

Any analysis has to be actionable. If not, they cannot be considered relevant as they cannot be acted on to change or solve business problems. Great analysis produces results that aids and points to specific solutions that one can take to solve the business problem. At the same time, the analysis must also come up with feasible solutions. A solution that requires such extensive changes to the business operation will unlikely to be accepted both by the business or the stakeholders.

Timely

The result from analysis must be timely. If the result is not timely, it cannot be available when there is a problem to be solved. Given the critical importance of timing, if a problem cannot be solved immediately in a timely manner, it will be better to focus on other problems that can be delivered on time

Comparison with BARCH

The G.R.E.A.T model is excellent for determining the success of an individual model. The model explains what the key elements are that make an Analytics model successful and accessible to

management. The model is also excellent in describing what a successful Analytics project should look like and it helps to guide analysts and managers to draft projects in an orderly manner and deliver them in user-friendly ways. There are several similarities to BARCH as well. The Guided principle is akin to the Business component of the BARCH model where they both seek to understand the business aspect. Relevant targets the business and also the cost and revenue aspect of business.

However, the model cannot handle more generic cases where there are less urgent problems. The G.R.E.A.T model requires models that solve urgent and immediate problems faced by organizations. That may not always be the case for the organization that operates in medium and long-term developments. BARCH focuses on the entire aspect of a project and can be extended to a more strategic level if required. BARCH, I would argue, is thus more versatile than the G.R.E.A.T model having elements that can deal with both immediate and long term requirements.

4. Framework Evaluation Summary

There are many comparison criteria that can be used to compare the models. During the discussion above, I have pointed out several limitations in the existing models and how BARCH addresses some of these. In this particular summarized comparison, I compare the various models in terms of the level of organization targeted, the focus of the model as well as the dependency on specific skills or knowledge.

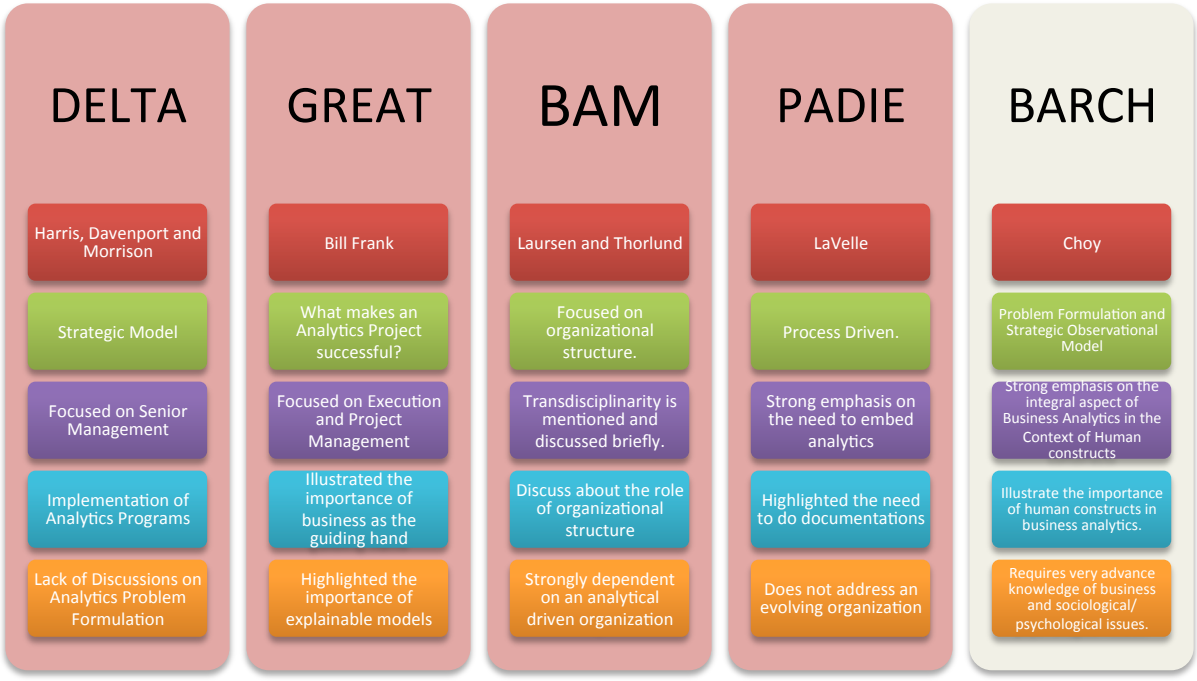


Fig 3. Comparison Framework

The DELTA framework is targeted at the senior management level and deals with strategic level issues. This is definitely a framework which helps senior management who wish to initiate corporate

restructuring to implement business Analytics as part of their road map. However, the framework does not address any business problems. Given that it is pitched at the senior management level, the model might not take into consideration the impact of the business Analytics solutions implemented. The DELTA framework also considers that the existing processes in the business are complete and not problematic. However, this might not be the case. Careful evaluation of the business is critical for the success of any business Analytics initiatives.

The GREAT framework is a framework that focuses on the project at hand and whether the project has the potential to bring positive change to the organization. The framework is targeted at the opportunities in the organization when they appear. The framework places emphasis on several critical success factors for projects. The framework targets project managers who need an execution plan for their Analytics project. Given the strong focus on execution, the framework suffers from a lack of consideration for problem formulation as well as impact analysis. However, the model recognizes the importance of business as the guiding force for successful implementations. The factors in the model are not within the control of the managers making it difficult for the framework to be applied beyond opportunistic scenarios.

BAM is perhaps the only framework that deals with organizational structures and how the structures affect Business Analytics. The strength of the structures also relates to its weakness. Given the need to formulate and solve problems, it is impossible for the model to address the critical needs of analysts. The framework is also highly organizational theory driven making it more theoretical than actually practicable. The model also describes the interdisciplinarity required for the structures to work which is a valid and important observation.

The only process driven model or framework is PADIE. The framework focuses on the processes in the organization, analyses them and then embeds the Analytics needed in those processes. The framework implicitly assumes that the model will explore the business and extract the key processes needed. The framework is suitable for analysts who can dive into the processes. However, there is a serious shortcoming in terms of the ability of the framework to cope with an evolving organization and business processes.

The BARCH model is more Analytics focused and driven by business needs. However, the model is not perfect. The model assumes deep understanding of the business, which might not be the case for analysts. The model is targeted at managers or analysts who needs to solve a business problem using Analytics. Even though senior management can use the model, it is not necessarily helpful to them in developing multi-year Analytics strategic plans.

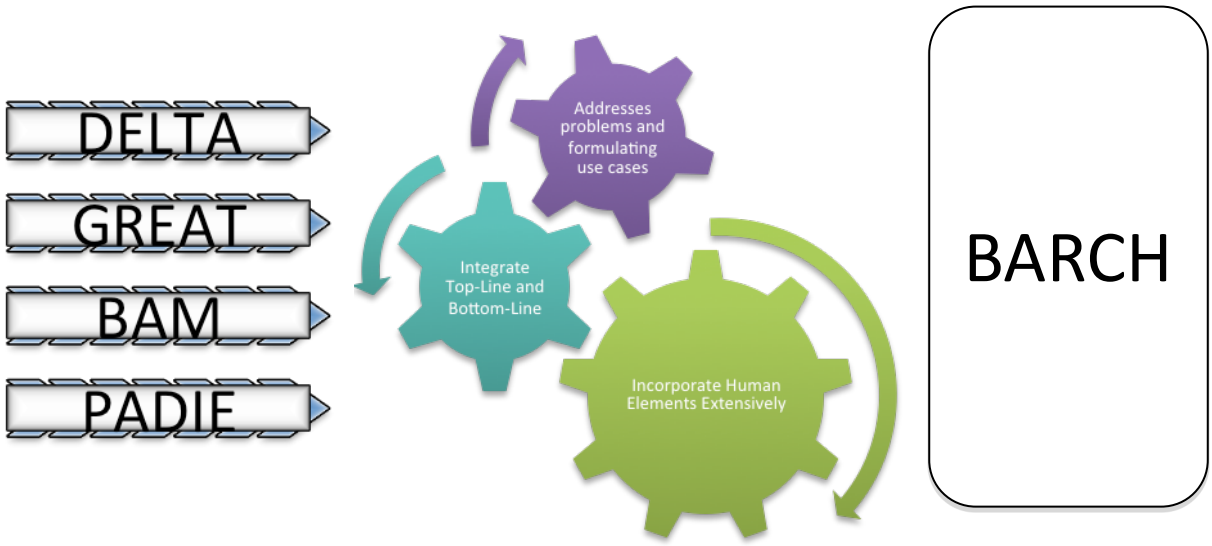


Fig 4. BARCH and its strength

5. Case Papers

Introduction

The majority of analytical models applied to many problems in the business environment address the problem superficially (Bose, 2009; Sara et. Al., 2010), that is, without understanding the impact on the business as a whole. Many Analytics projects have not delivered the promised impact because the models applied are overly complicated (Stubbs, 2013) to solve the root causes of the business problem. This problem is compounded by an increasing number of analysts applying Analytics to business problems without a proper understanding of the context, technique and environment (Stubbs, 2013).

Having experienced the difficulties of applying analytical models without understanding the context led me to believe there was a strong need to develop a simple and structured approach to managing business problems through the combination of both positivist and phenomenological research approaches. Given the huge impact Analytics can make to any organization, enabling business analysts to quickly understand the context of the problem and how to approach these problems in a structured manner helps to change the organization and deliver the impact of Analytics. This has been demonstrated in several case studies in literature.

One particular case Stubbs (2013) elaborated on the situation of an analyst who is given a particular task to identify opportunities that generate the largest amount of revenue. The analyst is self-trained using excel manuals and online learning materials. Using statistical models and drivers to analyse the cases, the analyst identified the key drivers that could potentially help the company to

earn millions of dollars in revenue. However, the model was plagued by two major issues identified in the literature. The first issue is that the model was built on a limited number of cases that cannot justify the predictive power of the model. The second issue is that the model's estimates have a huge confidence interval making every deal both a potential hit or a potential bust. This particular case highlighted several problems in the thinking process of the analyst that are not discussed in the literature that is dominated by positivistic discussions. From a phenomenological viewpoint, the analyst did not attempt to understand the business and the drivers behind the business and used only statistical drivers to drive business. At the same time, the analyst did not attempt to model the business process of the opportunities that ultimately determined the profit. From a positivistic stand, there is an utter lack of rigor in data collection and consideration of whether the data used can be compared across different opportunities. An example is the project budgetary considerations. A smaller budget in a smaller country might be better than a bigger budget in a bigger country even though the absolute amount might differ greatly. This is due to the size of the budget relative to the country's currency strength. The analyst also did not approach the problem from a business perspective. The key focus was on revenue that is just the top line without considering the bottom line (which in this case is defined as the cost).

One important aspect of phenomenology is the embedding of oneself in the process and understanding the interactions between oneself and the environment. The individual interacts with the surrounding to study both the surrounding without and with the individual's interactions. This helps the analyst understand how the entire business process works and, through the first-hand experience, they will be able to understand the subtle interactions with the business process. While the experience might not be universal, it does not in any way reduce the value of such experiences.

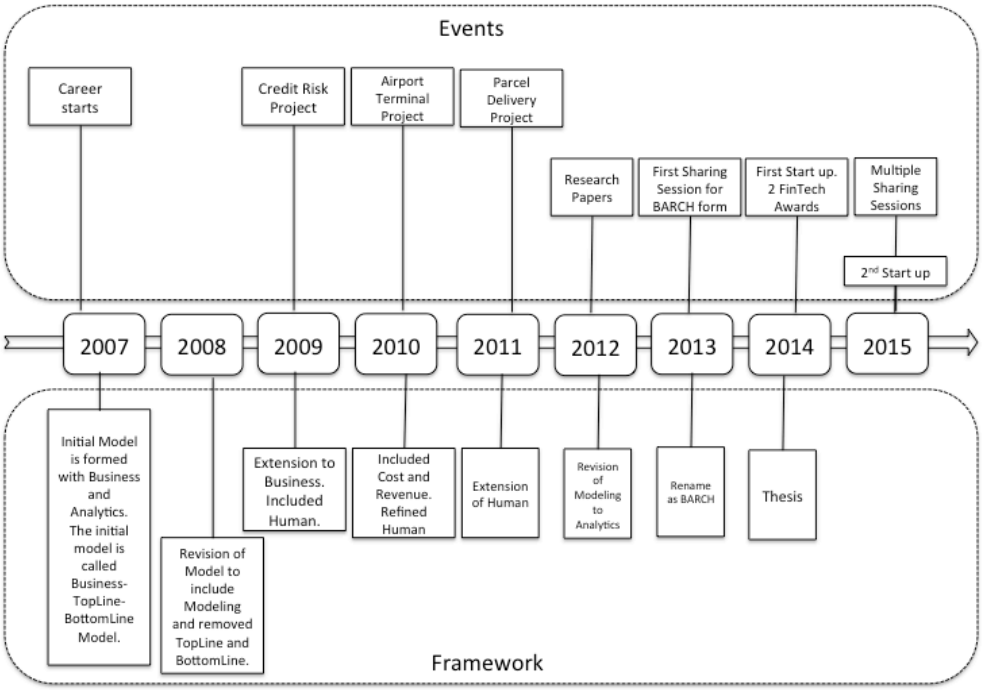
The training that the analyst receives from the book is merely a case study that does not attempt to put many thoughts into perspective and certainly does not mimic the reality sufficiently. Such attempts to bring textbook case studies to real life without sufficient consideration of the realities are misleading. Thus any framework proposed should always take 'real world' issues and complexities into consideration. If the analyst had been exposed to BARCH, the BARCH framework would have required the analyst to spend more time understanding the business and the human aspect of the business. By understanding the business, the analyst would have realized whether the drivers were sensible. Through Analytics, the analyst would have identified the shortcomings of the model and looked for alternatives. With two parts of BARCH, the complexities of modelling would have been obvious to the modeller and alerted the modeller to the dangers.

The incident also mirrors the Gettier problem mentioned in the earlier section on knowledge. Analysts might apply models with significant and positive results. This leads the analysts to believe it is their models that have impacted the business. However, this belief can be invalid and the true cause of the effect can be due to something unrelated to the model or the actions taken as a result of the model. Thus it is important for the analyst to be able to understand the effects of the models, and actions resulting from the models, which are linked to the results such as cost and revenue. An example is a prediction model for a marketing campaign. A sub group of customers is identified as high potential customers who are likely to buy something from you. To achieve better sales, you send them reminders or invitations. However, there is a major advertising campaign by a rival company that increases awareness of the products. These lead to comparisons online that result in

an increase in sales. Such events can lead to unjustified beliefs in the efficacy of the model. Thus it is important to use the model in multiple scenarios to establish its efficacy and the extent of its effect.

There was another incident that I encountered during my tenure as an assistant manager. We were engaged on a particular project for a bank. During the project, there were several analysts involved with the project and there was a fair amount of Analytics work required. The skill levels of the analysts fluctuate significantly and most have little experience applying analytical models to business problems. The analyst in question had some years of experience but lacked the domain knowledge of risk management in financial institutions. In one of her source code components, she programmed a particular construct to be a logical “equal” rather than “smaller than or equal to”. This major error caused a total reboot of the project as this piece of code is an input to the main analysis, which resulted in erroneous results and interpretations. The key factor that caused this was the analyst’s lack of knowledge of the business process that is compounded by the lack of framework for her to formulate the problem in a working form that can be solved properly or raise questions that will help the analyst to understand better. If the framework had been in place, the error would not have occurred. This problem is further compounded by the automation of modern analytical software, which prevents Analytics practitioner or users from detecting wrong until it is too late.

I formulated this framework based on 7 years of experience as an Analytics expert to companies and governmental organizations having been involved in hundreds of projects during this period. While the majority of business Analytics projects were successfully implemented, only a relatively small number of them had a positive impact on the organization (Stubbs, 2013). There are a variety of reasons for this lack of impact and one key reason is due to excessive imitation. Poor-quality imitators can result in significant negative impact on the perceived value of Analytics. Some companies and consultancies attempt to reduce Analytics to a simple push-button process or a standardized template that leads to customer mistrust when it fails to produce the expected results due to wrongful use. This failure is common to many organizations and can occur if they depend excessively on oversimplification or naive interpretation (Deloitte, 2012). Being a result driven person with strong focus on the impact, excellent deliveries of models is something that I expect to achieve.



Using the framework to evaluate three case studies across different industries, I plan to demonstrate the effectiveness of the framework dealing with Business Analytics problems. Through the case studies, I hope to emphasize the importance of contextualizing the problem in a business framework using both positivist and phenomenological approaches. The three cases were selected based on their direct impact to the business and the outcome of the solution implemented. To demonstrate influence, it is critical that the work is appreciated by the organization that commissioned the work and the organization has taken actions as a result of the findings. In all three cases, there were impacts on the business either in the form of a transformation of the business model or a global level implementation. The case studies also highlight the importance of reducing the problem into smaller individual units of analysi which can be linked up in a complementary way to answer the bigger problem. I hope to demonstrate the importance of my learning from this critique of my existing works through its applicability to business problems, the efficacy of reducing complex problems into simple solvable units and form meaningful results which will be the basis of action for businesses and organizations. Additional case studies are included to highlight other cases where BARCH was applied to solve Business Analytics problems which led to some form of indirect impact, or impacts that cannot be quantified. The co-authors for the papers were either reviewers or by courtesy.

5.1 Credit Risk Scorecard (Choy and Ma[10], 2011)

Credit risk scorecards are important tools in the tool box of the banking industry. The concept has been widely used to control consumer credit risk and has been extended to small business credit risk (Anderson, 2005; Choy and Ma, 2011;Thomas et. al. 2002). The earliest credit scorecards were

[10]Ma is the co-author as she was the reviewer and thus included as part of protocols.

developed by Credit Scoring Consultancies as a way for finance companies to identify risky customers that should not have been given a loan. Due to their proprietary nature (or aptly statistical nature) (Anderson, 2005), few could understand or comprehend the underlying mechanism of the scorecard at that time. Early practitioners of credit risk scorecard modelling spent massive amounts of time refining the techniques used to build the scorecards. Besides refining the techniques, they spent a lot of time explaining the mechanism and philosophical approach to the finance companies to convince them to use the tool. This is something that is also seen in the BARCH model. The early practitioners understood the importance of the business process and they actively explained the relevance of the model to the business. This demonstrated their thorough understanding of the business which enhanced their credibility as evidenced by the prevalence of credit scorecards. The refinement of the models is a classic example of the interactions between business, Analytics and human in the BARCH model. Through their interactions with users and business, the practitioners actively transform the type of the models used to explain and predict better for the users.

As time went on, more and more people gained an understanding of the mechanism of the credit scorecard and were more willing to adopt the model to manage their business. The sudden rise in the consumer credit market directly led to the rise of the credit scorecard industry marking a new milestone in the industry (Lewis, 1992). Many big credit scorecard consultancies were established during this period of expansion such as FICO and Experian that resulted in the significant disparity in the approaches taken to quantify the risk. The main criticism against credit scorecards then was that the variables bear little relation to the variables modelling them and that the definition used in the modelling can be rather haphazard and offers little help to finance companies who are trying to manage these risks. This strong opposition is also voiced by some authors (Capon, 1976; Rosenberg et. al., 1994). While there has been much refinement of the credit scoring techniques in the banking world (Eisenbeis, 1977; Eisenbeis, 1978), many criticisms have not been satisfactorily resolved. The often cited misuse of variables and information to quantify risk can be seen as the breaking down of the relations between business and Analytics. This can be due to at least two issues. The first issue is the need for innovation. Very often, innovation in the market seems to be the product of repackaging than true innovations and these actions led to more damage than good. Another major issue is the capabilities of the analyst. Unlike the early practitioners, they lacked the connection to the business which resulted in their inability to comprehend the various aspects of the business. This inability to understand is one key reason why they use variables which does not relate directly with the business. BARCH model alleviates this problem by ensuring the user has to refer to the business as much as possible when they use it to solve a problem. The reference to the business also ensures that the Analytics model applied has to be relevant.

With the advent of Basel II, there has been widespread discussion about the definition of a bad debt account in the context of a credit portfolio. The accepted definition for Basel II is any accounts with a history of ever been more than 90 days past paymentdue date within a period of 12 months is considered to be a bad debt account. This definition is controversial as different financial products behave differently. Some credit products such as mortgages take a long time forany accounts to satisfy the bad debtdefinition while in other cases, the period is too long and most accounts will be considered bad debt by then (Thomas, 2002; Siddiqi, 2006). Thus a proper definition is critical to both the proper management of risk as well as the operational needs of the banks. The key problem with Basel II and III is that the definition of bad is derived primarily from 2 sources. The first source

comes from regulatory agencies who are not directly managing the risk and the second source is primarily from academics who are not practitioners as well. This results in poor tie in with the business.

Credit risk scorecards are designed to measure the probability of an event happening. To be able to measure such events, one must define the event in a manner that is easy to measure and is not confused with other events due to combination of the events. Typically, the target scenario is something that is easy to comprehend and precise. The earliest credit scorecards used simple target events such as predicting whether a customer will develop a situation where his credit history will indicate that the card has a record of being more than 30 days past payment due date within the next six months. The simple definition is crucial as it helps to tie in to an aspect of the business. The gradual improvement in raw computing power has resulted in ease of building more complicated models that attempt to capture more variations of the bad than what is traditionally used in modelling. However, the complicated measures reduce the relevance of the measures to the business. They also further complicate the Analytics model. Complex measures will require more sophisticated models which are not necessarily better and can increase the amount of work needed to comprehend and implement them especially models such as neural network which require a lot of resources.

Using the BARCH model to understand this problem, it helped me to formulate the problem from a different perspective which led to a new solution. BARCH directly focuses on the missing gap between the business and the definitions used and constructed by the stakeholders. By linking the various aspects, I examine the various options and models taking into considerations the factors in BARCH to develop a Markov Chain solution. The solution also demonstrated the simplicity of the Markov chain approach that solved both problems simultaneously. The bank eventually adopted the approach to solve their operational needs while maintaining Basel II requirements.

Solving the problem requires the analyst to understand the business of credit lending. The concept of credit and lending on interest is not a modern phenomenon and has existed for thousands of years. The basic principles controlling when credit is granted and the types of credit available have not changed much since Babylonian times. Credit has been defined largely by the prevailing moral climate and the resulting legislation to address the problems resulting from its use and misuse throughout history. Much of the historical record that relates to the laws and restrictions on lending and the operation of legalized credit markets is well defined.

The technological advances have changed the way in which credit is marketed, granted and managed. These changes facilitated new mediums of credit such as mail order and credit cards which further fuel the desire of the consumer to acquire larger numbers of the ever increasing range of consumer durables on offer. Limitations on interest rates have existed throughout most of history with relatively few exceptions.

Credit comes in a variety of forms that are characterized by six key characteristics:

- Whether credit is given under a secured or unsecured term. A secured credit is one where there are specific items or collaterals named in the agreement that can be claimed if the terms of the agreement are breached.

- Whether credit is amortizing or balloon in nature. An amortizing loan is one that is repaid slowly and gradually over the duration of the agreement. A balloon loan is paid off as a lump sum at the end. In some circumstances, the loan is a mixture of both amortizing and balloon loan.
- Whether the credit agreement is fixed sum or running account. A fixed sum credit account is one that disburses only a fixed amount of money. Running accounts basically gives loans up to a certain amount that the user can use up to. Any remaining amount in the loan is the difference between the approved amount and the balance which is the amount borrowed. Once the balance is paid off, the loan amount is available again.
- Whether the purpose for which credit is obtained is restricted or unrestricted. If credit is restricted, then whether credit is provided on a credit sale or conditional sale basis. There are several variations depending on the market. An example is the housing bridging loan which is contingent on a full mortgage that has to be taken up.
- Whether there is a two parties (creditor-debtor) or three parties (creditor-debtor-supplier) relationship. The first relationship is more traditional and common in the industry. The second relationship is more common in the funding and credit marketplace where people can engage in lending behavior.
- The cost of borrowing.

The annual percentage rate (APR) is a standardized way of representing mandatory charges that are applied to a credit agreement. It is critical to note that the APR is not the same as the interest rate charged by a lender. The APR represents the total cost of credit expressed in the form of an interest rate. This includes interest and other mandatory charges such as arrangement fees, valuation charges, product guarantees and brokers' fees. Optional charges are not included in the APR calculation. The main types of credit that individuals can obtain are:

- Mortgages
- Personal loans
- Retail loans
- Hire-purchase agreements
- Card accounts
- Charge accounts
- Revolving loans
- Mail order accounts
- Payday loans
- Pawn loans
- The different organizations that provide credit include:
- Banks
- Building societies
- Finance houses
- The government
- Pawnbrokers
- Licensed moneylenders

Banks, building societies and finance houses supply the vast majority of credit in the markets. Government, pawnbrokers and moneylenders represent only a very small sector of the market which is one that tends to be utilized by people having difficulty securing credit on standard terms from mainstream providers. However, in smaller markets or special economies, government can be a major supplier of credit.

Most lenders will attempt to predict the likelihood of a potential customer defaulting on their credit and then make lending decisions using the prediction. The good and bad likelihoods act as approximations to profit and loss respectively. Usually, the measures correlate closely with profitability. However, every lender has their own overheads and operates with different levels of profitability. Thus a customer deemed creditworthy by one lender may be not be the case by another.

For some time, lending decisions were made entirely on the basis of an underwriter's subjective assessment of an individual's creditworthiness which relies on the underwriter's experience. Today, judgmental decision-making is the exception, with most lending decisions based on credit scoring which is the application of mathematically derived forecasts of future repayment behaviour through statistical modelling. While the approach is not perfect, it has demonstrated a number of benefits over the judgmental approach:

- It provides a more accurate assessment of risk.
- The automated decision systems incorporating credit scoring has led to faster and more efficient processing of credit applications through a centralized control function.
- Credit scoring is consistent and replicable.
- Credit scoring does not tend to display the unjustified prejudices that human underwriters sometimes expressed against certain sections of society.

Below are some examples of good definitions of 'bad' accounts and contrasting them with complicated and infeasible definitions (Choy and Ma, 2011).

Bad Definition for Modeling (Choy and Ma, 2011**)**	
Good	**Bad**
Ever X+ DPD in 3 Months	2 Times X+ DPD in 4 Months
Ever 30+ DPD in 6 Months	6 Times 30+ DPD in 12 Months
Ever 60+ DPD in 9 Months	2 Times 30+ DPD and 4 Times X+ DPD in 10 Months
Ever 90+ DPD in 12 Months	2 Times Consecutive 30+ DPD in 12 Months

Table 1: Bad definitions

The problem with more complicated bad definition is the difficulty in truly understanding the outcome. Let us contrast the good and bad definitions and use the row 3 definitions from table 1. If you were to ask an analyst what it takes to be a bad customer, the answer will be the definition and

you wonder, what about customers who are 1 times 90+ DPD or 3 times 30+ DPD in 10 months? Another possible situation that might arise from this definition is the simplification of the complex definition. The first condition is an extension to the second condition which implies that we can simplify slightly to '2 Times 30+ DPD and 2 Times X+ DPD in 10 Months'. One severe issue with using this type of definition is the time period needed. Given 6 times delinquent in 10 months, the probability of such an event will be very unlikely, resulting a small target population for modelling (Choy and Ma, 2011).

The bad definition is critical to the business of credit lending as it forms the bedrock of the entire process. The bad definition will define the good and bad population and the way the model predicts the defaulters. However, because of regulations especially Basel II, certain bad definitions were assumed for the sake of simplicity as well as compliance with the regulations. The approach yielded models that are difficult to use on certain portfolios and does not meet the needs of risk management and operations. To handle the issue, anew and appropriate Analytics approach has to be formulated. The new approach has to be consistent with both the regulatory requirements as well as the banks' internal needs. The simple application of several properties of Markov Chain is the formulation approach that can unite and solve the common problem.

Markov Chains, also known as transition matrices, are mathematical models which define the probability of an object moving from one state to other states. Depending on the data available, there are several ways to build such a matrix. Below is the mathematical form of the matrix.

States	**A1**	**A2**	**. . .**	**A(N-1)**	**A(N)**
A(1)	P(1,1)	P(1,2)	. . .	P(1,N-1)	P(1,N)
A(2)	P(2,1)	P(2,2)	. . .	P(2,N-1)	P(2,N)
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
A(N-1)	P(N-1,1)	P(N-1,2)	. . .	P(N-1,N-1)	P(N-1,N)
A(N)	P(N,1)	P(N,2)	. . .	P(N,N-1)	P(N,N)

Chart 4:Hypothetical Transition Matrix

Each entry in the matrix represents the probability that an object will move to this state given that it starts from the state on the left per turn (usually defined as the time to transit which in this case is one month.). Total sum for each will be 1 for closed systems. One of the interesting properties of the Markov chain is that one could calculate the average time spent in each transition states. This calculation is only possible in cases where the matrix contains only transient states (Referring to the case where the row summation does not total to 1). Because of this property, it happens to be

uniquely qualified to solve the problem faced in solving the performance period and delinquency to default values.

Let us consider a matrix Q where the states are numbered T = {1,2,…,t} as the set of transient states.

For each transient state i and j, let m_{ij} denote the expected total number of time periods spent in state j given the starting state of i. Reorganizing the formula yields the following result.

Where = 1 when *i = j* and 0 otherwise. Let M be the matrix containing m_{ij}.

Converting it into the matrix form yields the following equation

$$M = I+QM$$

which can be transformed into

$$(I-Q)M = I$$

and with a little tweak becomes

$$M = (I-Q)^{-1}$$

The model must work in tandem with the operation of the credit lending process. The first important component is revenue.

Credit providers generate revenue from a number of sources, which include:

- Arrangement and annual fees.
- Interest.
- Interchange fees.
- Insurance products which include:
- Income protection.
- Critical illness cover.
- Payment protection.
- Card protection.
- Mortgage indemnity insurance.
- Late fees and penalty charges.

While interest is perhaps the cost foremost in the minds of borrowers, for many credit providers a significant part of their profits comes from other sources, particularly the selling of insurance and the charging of late fees and penalty charges. These additional charges tend to feature less in advertising material than interest rates, and are generally given little attention by consumers.

Therefore, many lenders see increasing existing charges or introducing new ones as a better way to increase revenues than increasing interest rates. Besides revenue, we also need to examine the cost of the process.

The major costs incurred by credit providers in supplying their products are:

- Cost of funds.
- Bad debt (and provision).
- Fraud.
- Promotion and advertising, including the costs of incentives such as loyalty points, interest-free periods and discount offers. This also includes targeting existing customers to retain their business and thus reduce the incidence of attrition.
- Application processing.
- Infrastructure costs comprising, capital expenditure, depreciation and day-to-day running costs. Of these costs, the two most significant by far are the cost of funds and the cost of bad debt. Between them they can account for more than 70 percent of the total cost of credit provision for some credit products.

The Markov Chain model is simple to apply and can be derived easily from the reports commonly used in the bank without requiring additional analysis or work. This makes it easy to compute and for analysts to use. At the same time, the formal derivation strengthened the human's faith in the accuracy of the model. During implementation, there are resistances to the model but through careful explanation and good demonstration of the efficacy of the model, it was accepted.

Learnings

The BARCH process was applied in this project to deliver a model that makes a difference to the operation of the banks. The Basel II model is difficult to apply without many alterations. The approach was endorsed by an international bank and a regional bank to facilitate their risk management needs. Similar concerns raised in the paper were subsequently raised in the BASEL 4 consultation papers especially on the issue of non-unified computation of risk period and definition for different banks in IRB (Internal rating) models (BIS, 2015). To aid in the process of credit risk management, BARCH distilled the business problem into a form that directly linked the business with the model as well as the expected output. The distillation process involves careful study of the lending business and how the regulatory environment has developed problems for the business. This is supplemented by the study of the revenue and cost structure of the lending business. By understanding the gap in the situation as well as the positions of the stakeholders in the project, I begin to do research into the types of model that can be used to solve the problem. This led to the Markov Chain solution which is simple and elegant. The process also allows me to refine the BARCH model and made the BARCH model more relevant and easier to use.

In the process of solving the business problem, I came to the realization of the importance of the business model and the interactions with the problem. All business problems ultimately come about because of the business and its processes and attempts to solve the problem alone without

considering the business is a level of abstraction that nullifies the solution. The business model understanding cannot be superficial. The understanding has to be deep in the business and also the human interaction aspect.

This directly influenced the creation of the BARCH model especially the business and human aspects that are key components of the learning from this project. The business problem is the result of the business and business process. As the problem is derived from business, we have to first understand the business and the processes. The business is defined as an entity and the process structure the entity's operation. In the project, I found the problems of the disconnect between the regulators and the banks. To remedy the situation, the most direct way to solve to the problem is to produce a business framework that creates a common ground for both the regulators and banks to work together. However, the frameworks in the literature are focused on the description of the businesses and how the components of the business interact with one another. This does not serve the purpose well. To tailor to the needs of BARCH, we redefined business' role in the model and connected the business and the objective of the problem.

Even though we connected the problem to the business, it is important to also link the problem and business to the human aspect of the business. As with the phenomenological approach, the problems are always interacting with the human aspect of the business be it the banks or regulators. This link up is crucial, as any models applied to the business will impact human. By including this consideration, we will be able to address human problems earlier and understand how the model will impact the business more accurately as well as the long-term impact. This experience confirmed my long held belief that a good understanding of the business is critical to the modeling process and ensuring the results are relevant. Interestingly, this view is echoed in a variation by Warren Buffett who states that one should not be involved in a business until he understand the business thoroughly.

Another important learning from this process is the importance of the level of abstraction. Humans are not too fond of excessive generalization and abstraction, and are not entirely able to grasp the complexities of many events. Thus the inclusion of the human aspect assists in making the models comprehendible to humans. By including the business and human elements, the results make better sense in the context and provide greater insights and benefits. This was my first affirmation that BARCH was along the right lines, that it is a more thorough understanding of the business that is needed to solve the problems through Analytics.

5.2 Airport Terminal Logistics (Choy, Ma and Cheong[11], 2012)

The airport terminal operator in this case is a major player in the region and operates one of the world's most highly rated airports. The company is involved in the development and management of several projects in the regions. The company has been extremely successful in its ventures and has won much acclaim for the airports that it manages. Through the years, the company has faced increasingly stiff competition from other rivals in the region. To maintain its leadership position, it

[11]Ma and Cheong are the co-authors as they are the reviewer/sponsor and thus included as part of protocols.

turned to Analytics in an attempt to improve the customer satisfaction as well as the overall profitability of the business. They needed insights to provide them with ways to make their business more sustainable with the primary aim to increase revenue. The airport terminal business occupies a niche environment that meant that the feasibility of an Analytics project required a close examination of the many areas of the airport that combine together to constitute an airport terminal. Both DELTA and BARCH provides interesting insights to this problem.

The global air travel market has been growing at a steady rate for the past 20 years. Since 2001, the total number of air travel passengers has increased by 25%. This increase translates to approximately 2.25% growth annually in passenger load. With increasing passenger numbers for air travel, the common expectation will be that the revenues would increase at similar rates. Unfortunately, expenses have also increased at approximately similar rate of 80% that effectively wipes out the profit margin for most airlines. As a result of the combination of poor profit margin and unprecedented high competition between various airport terminals, the environment is very hostile and difficult for airport terminals to increase the existing fees that they charge airlines for the use of terminal and services without rendering additional services.

The main challenge and objective to airport terminal operators is improving the profitability of the terminal business. Most airport operators generate revenue via one of the following 3 options.

- Airport gate and parking rental
- Check in counter space rental
- Consumer space rental/sales

The first two sources are highly dependent on the airlines. Thus increasing income from these sources is unlikely if not impossible as most of the airlines are making a loss or minimal profit due to high expenses. Thus from a business perspective, both sources are effectively dead ends. That leaves the third source as the most preferred source of revenue generation.

Passenger satisfaction has become a critical and key performance indicator for the survival of the airport operations. Famous and critically acclaimed airports such as Incheon, Heathrow and Narita airports have set up customer satisfaction surveys and improvement schemes to improve their reputations and identify areas of shorting coming. Even regional airports and smaller airports are beginning to change and evolve to engage customers in the retail area. To be a passenger focused business, the airport has to maintain excellent scores in customer satisfaction both in perception and reality. This helps to ensure that passengers are more likely to spend more time and money in the premises which translate into improved profit.

In order to increase revenue from this avenue, the best and easiest way is to increase the traffic flow through the consumer space. With the fixed layout of the airport that is mandated and controlled by regulations, this poses a major challenge for any airport with limited space for non-essential services. The alternative is to increase the overall passenger population handled by the terminals which directly translates into higher traffic intensity. To increase overall population handled by the terminal, this means an increase in the number of flights which is constrained by the airport infrastructure. Given such restrictions, the operator that needed help in this case was ready to give up. The DELTA framework could offer no solution to the problem. The organization has the data that it needs, the enterprise infrastructure, the leadership to push for it, clear targets to aim for and the

right analyst. However, the framework cannot give any suggestions on the areas to work on. However, BARCH presents an interesting alternative. The BARCH framework first examined the requirements of operating an airport terminal.

The responsibilities for airport terminal management vary broadly and are based on the physical requirements as well as the needs and interests at each particular airport. Such variations are due to the result of local culture, geographical locations and traveller patterns. For example, airports with a lot of transit passengers are more likely to have stores and shops that provide travel essentials than branded goods and tobacco. Most management include key essential services which are grouped under terminal facilities maintenance, base building electronic systems and airline operations electronic systems.

Terminal Facilities Maintenance as defined in the context of airport terminals covers trade services that are concerned with the maintenance and repairs to terminal and concourse facilities, which is dependent on the location. Base Building Electronic Systems covers telephone systems, fire alarm, security access control and other related information technology systems. Typically, most electronic systems will fall under this area. Airline Operations Electronic Systems covers flight information displays, baggage information displays, ramp information displays, gate information displays and other related technology, depending on the location.

Most management companies have specialized staff dedicated to the day- to-day operation and administrative management of their company. These staffs have specialized knowledge and skills which are not commonly found and are highly in demand. Under most circumstances, the company will directly employ the dedicated operational and administrative personnel. In certain jurisdictions, the management company has to cooperate with a local company to provide the personnel. This is commonly done when the airport terminal management company is a foreign country owned entity running national airports for another country. Private firms that have been contracted by the company may also employ the dedicated operational and administrative staffs that counts towards the manpower of the management company. Because of the niche area of operation and complexities of regulations, it is common for the business to look into various ways to optimize the business. The high levels of specialization also imply that the business requires very specialized knowledge that cannot be accessed or mastered easily. This is something very unique. The business is tightly linked to the business users and they are the central actors.

The duties and tasks of the terminal management company may be grouped into several categories including:

- Liaison activities
- Contract administration
- Corporate administration
- Financial administration, and
- Operations coordination and administration.

In the industry, performance and maintenance data are recorded using computerized maintenance management systems operated by the specialist staff or selected vendors with expertise in the area and system. The capturing of the information is critical for the success of the company as they measure the various aspect of the business. This provides information on both the financial status of

the company as well as the processes of the business. The type and frequency of performance reports produced by the companies in the industry varies considerably. Most companies publish no regular performance reports unless requested by airport staff. This usually results in an opaque environment where management does not have an overview of the situation. In all cases, airport staffs have the ability to make inspections, request for reports and recommendations as needed.

Any failure to perform the obligations under the agreements between the companies and the airports is commonly considered to be an event of default. Such an event permits and grants the airport the right to perform and relieve the duties of the management company which effectively fires the company. Given the enormity and complexities of managing the terminals, both parties involved want the cooperation to be successful for the airlines, terminal management and the airport. Thus they are not really interested in applying any penalties except under extreme circumstances.

Financial penalties are rare due to poor performance for airport management companies. The unique nature of each airport makes performance comparisons between airports difficult or at times highly distorted. It is difficult to compare results when studying the operating costs and costs per enplaned passenger for airports, as a result of the structural differences, the nature of the passenger traffic, and the costs included or excluded at each airport. Typically, the performance measures are established through a process of comparing airport with similar characteristics. Once similar airports are identified, the management companies are then compared by checking whether the scope of responsibility is similar. When similar comparisons are made, the cost per enplaned passenger has clearer differentiation. The cost and revenue can be seen as driven by the number of enplaned passengers that is the result of the completely different operating environments.

Most airport management companies calculate their rates and charges for costs that flow through the operations and must be collected directly from the airport. Each management company has its own adopted unique rates and charges models and methodologies that rely on a number of cost centres. The cost allocation methodologies adopted by each management company best suit the specific business requirements and caters for equitable allocation of costs to the participating airlines at their location. This cost is usually calculated at the passenger level and thus we should focus the analysis on the customers. From the business analysis, we understand the complex cost structures and how the performances of the airports are being compared. The DELTA framework offers little if any insights into these matters. The DELTA framework does not in any way address the target or how targets are constructed. Leadership is hardly a major factor here given the niche and specialization of the roles. In contrast, BARCH focuses on the business with relation to the human factors. The business is also viewed with revenue and cost involved.

The terminal management company and operator that is in need of help has 3 terminals under their management in this region. From past experiences with time series data, it is expected to observe some form of seasonality effect of the passenger throughout the year. To great surprise, there were no major seasonal effects in the weekly total number of passengers from Figure 1. The three different terminals have curves ranging from 200,000 for the smallest terminal to 280,000 to the most congested terminal with minimal variations. The passenger load for terminal 1 is also higher than the other terminals when we break down the total number of passengers by days of the week in Figure 2.

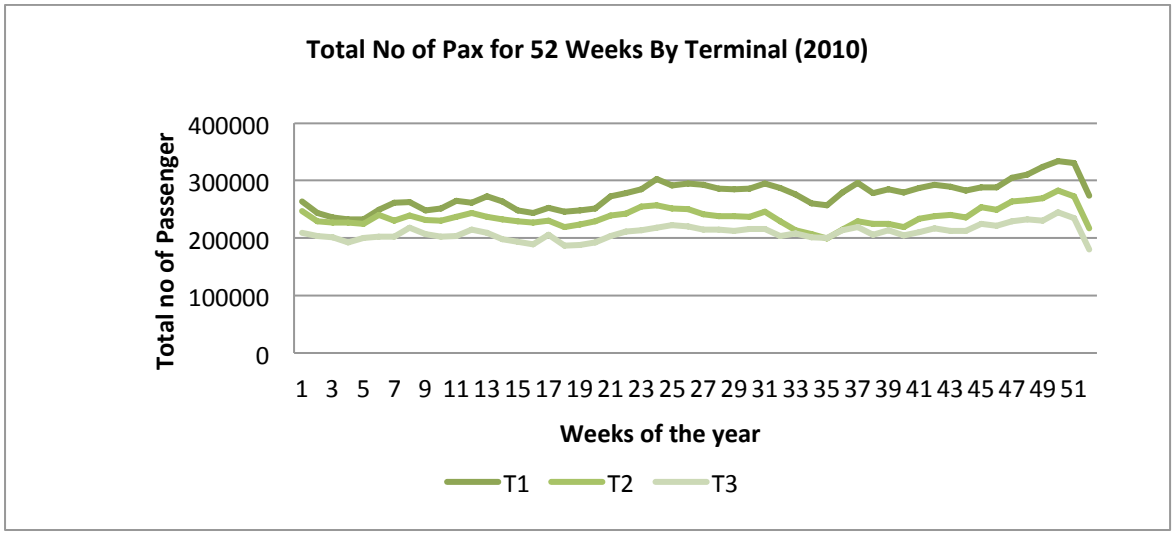


Figure 1: Total No of Passengers for 52 weeks by Terminals

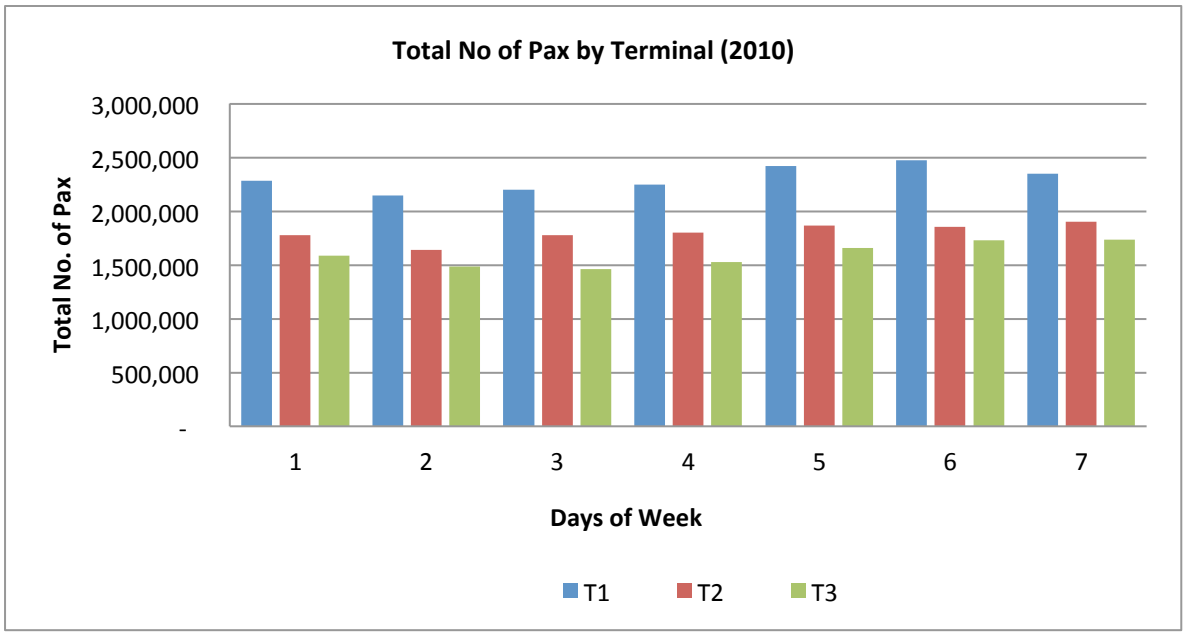


Figure 2: Total Passenger Load by Terminal With respect to Days of week

Close examination of the overall flight load of each terminal for each day of the week for the entire airport point us to any load imbalance given that the airport terminals are designed to be the same in capacity (Choy, Ma and Cheong, 2012). From Figure 3, it is clear and obvious that the impact of overloading of flights in terminal 1 even though the number of passengers handled is relatively close for the different terminals. To better understand and assess the discrepancy in the report, we have further analysed the passenger per flight ratio. From Figure 4, we notice that the passenger per flight ratio for terminal 3 is almost double that of terminal 1.

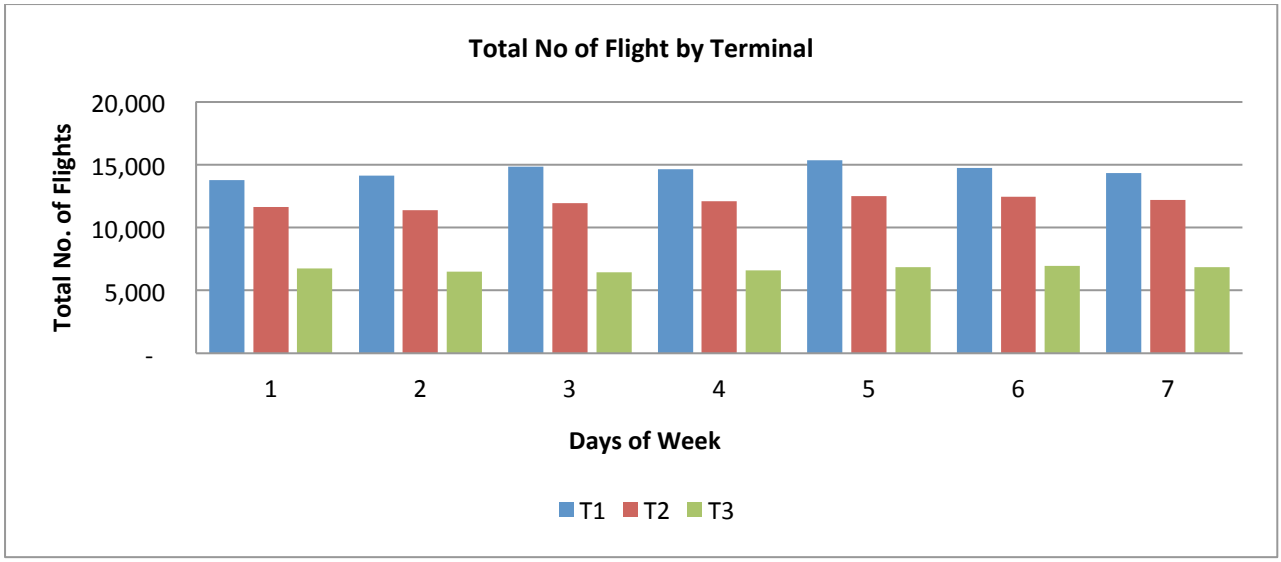


Figure 3: Total Flight Load by Terminal With respect to Days of week

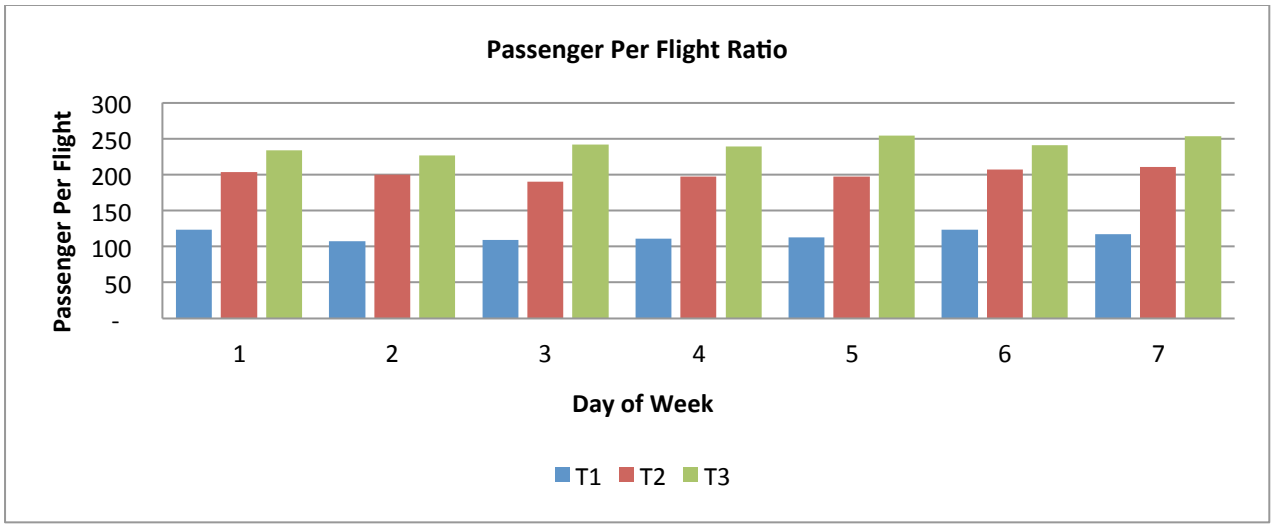


Figure 4: Passenger Per Flight Ratio With respect to Days of week

Given the similar level of capacity each terminal has to handle the flight loads, the disparity in the flight load handled is unusual and merits further investigation. Such discoveries create new opportunities for the business to do something innovative. The difference in the passenger per flight ratio indicates some operational differences between the types of aircraft or route that these terminals handled. Possible reasons raised by the airport management include bigger aircraft such as Airbus A380 as well as heavy traffic routes to top major cities. Through discussion and engagement with the management and the business users, we were able to zoom into the possible factors and reasons for the discrepancies. In this particular engagement, there were further reinforcement of the importance of business in BARCH and how it relates to Analytics and Human. With the new leads and reasons, we then try to identify the issues and propose solutions that can be used to rectify the bottleneck and improve the situation.

From the analysis, we have identified several interesting issues with the airline assignment to the terminal.

- Unbalanced flight load utilization for each terminal
- High variation in the passenger per flight load.

Unbalancing of workload at different terminals at the airport means that the resources such as gate or check-in counters are either over-utilized or underutilized. Overutilization and underutilization both creates scenario that reduces revenues and increase costs which is undesirable. Flights need the gates in the terminal to load and unload passengers. If the number of flights at the terminal is overloaded, it means that when the flights arrive, the arriving airplane may need to wait a long time on taxiway to occupy an available gate. Departing passengers in the congested terminal may also suffer longer waiting time at the check-in counters that leads to lower level of customer satisfaction. Due to the prolong waiting time at the check-in counters, the passengers may not even have enough time to shop at the retail shops before boarding. On the contrary, another terminal's resources are not fully utilized to their maximum capacity and wastage may occur. This is a serious issue for the airport operators to resolve so that the overall profitability and customer satisfaction will be improved. Most papers in the literature focused on optimizing either the flight load or the passenger load. The current issue is the outcome of optimization of flight load and passenger load independently for each terminal. Such problems in Analytics are optimization problems

Optimization problems which involve assignment issues are modelled as Integer-Programming (IP) models. The main objective of the problem is to balance both the flight load and passenger load simultaneously. This problem is not an isolated case which can only be applied to airport terminal operations. The approach can also be applied into various logistic problems where the load in different warehouses are unevenly distribution causing some service level failure. It is also a similar problem for container port operations where the management needs to apply some rules to assign the shipping lines to different container terminals.

The proposed model has been preliminarily tested using publicly available data online which include the flight in and out of airport for the whole year, airline-terminal current assignment and maximum capacity of the plane based on aircraft type and specific load factor for the airline. The problem has been solved using the SAS/OR optimization software. The problem is solved within 5 minutes for a realistic problem size of 87 airlines with 232,000 flights and nearly 40,000,000 passengers.

With the new assignment, we can see that the number of flights handled for each terminal is more evenly distributed with respect to their capacity. The new assignment has also changed the distribution of the passenger to the terminals based on the capacity load. Previously, terminal 1 heldthe largest passenger load even though it is not the biggest terminal. The new assignment has made terminal 2 the main handler for the passenger loads shown in Table 1 and Table 2.

Terminal	Before	After	Capacity
1	101,702	68,811	105,850
2	84,187	89,644	127,750

3	46,979	74,413	102,200

Table 1: Terminal Flight Capacity Comparison (Annualized)

Terminal	Before	After	Capacity
1	16,113,471	10,525,765	22,000,000
2	12,619,233	18,380,060	27,000,000
3	11,183,393	11,010,272	21,000,000

Table 2: Terminal Passenger Capacity Comparison (Annualized)

A comparison of the flight load indicated that the workload is more evenly distributed with respect to their capacity compared to the past. This indicated possible opportunities to increase the number of flights or airlines using the terminals to increase potential revenue while maintaining the flight load and passenger load that result in acceptable customer satisfaction levels.

Learnings

To solve the terminal management company's problem of increasing the revenue while subjected to constraints, the entire problem is framed up using the BARCH model. This proved particularly challenging. Even though the users are in need of help and solutions, they came with a mind-set that they understand the business very well and that they have explored the entire universe of solutions. They were adamant that most people do not understand their business and that bulk of the proposed solutions were dismissed. This is despite the fact that sufficient amount of research has been done to understand the structure of their business. At this point, it is clear and apparent that the management is operating purely from a business aspect without considering the human factor. The management is overly concerned with the need to increase revenue and decrease cost to the extent of ignoring the need to make slight changes to the business. In the entire process, even though the leadership is clearly in favour of Analytics, their actions are detrimental to Analytics and highly counterproductive. According to the DELTA framework, the determined leadership will have been a major boost to the Analytics initiative. This is coupled with the necessary data, enterprise wide systems, targets that are objective and analysts who are trained. What they missed is the importance of human factors. This learning makes me review the position of human factors in the model and strengthen the need to take the factor seriously.

The project also highlighted a couple of serious deficiencies with the DELTA framework. As mentioned above, even with the various components in place, it does not guarantee that Analytics will be given room to flourish. It is very crucial to understand the business as well. For a niche business like airport terminal management, the kind of business they operate behaves very different from other business. The business is highly regulated and has strong dependencies on visitors and airlines. By analysing the business, we can see that the cost structure of the airport terminal management is mostly fixed especially given the sunk costs of equipment, limited space and niche labour market. All these factors result in a static business model. Solving static business models require some innovation in thinking about the business. One key learning from the project is that the

business model cannot change for an industry with a lot of fixed cost. To improve the business, one has to focus on the operation of the business and adjust it. In this case, all that is required is to move and shift airlines between terminals to make it feasible to get new airlines into the airport and distribute the passenger across the terminal to improve passenger experience. The solution does not require a change in business model and no additional cost. This also highlights the need to link the business and the cost of the BARCH model.

Unlike the previous case, we did not follow the order of BARCH in a linear manner and several components were discussed together with the other components. Such discussions are more useful as they weave the intricate relation between the cost, revenue and business. In this case, the analysis of the combination of business and cost yielded insights of the restrictions on the business and how the Analytics model needs to incorporate this restriction. This project highlighted the problems of handling human and their influence on the success and failure of any Analytics projects.

During this project, I further developed the BARCH by modifying the Business aspect and focusing on the most relevant business processes. Unlike the earlier problem and business, this business is larger and more complex which posed a number of interesting challenges that required refinements to the BARCH model to be made. The manner of deployment was made by involving the management staff and end business users in the discussions. This part of the project brings out the criticality of the human factors. Without the phenomenological analysis of the human constructs of both the initial team and the senior management, we would not have been able to come up with the solution. The human aspect of the problem is the problem of the people being too steeped in their beliefs about their understanding of the business and refusing to listen to a new perspective.

This ties the business and the management together. Through the modelling process, we incorporate the cost and revenue processes to link them up and ensure that the model we proposed does not impact or change the cost structure significantly. These measures determine the survivability of any business. Unlike the management's attempt at defining KPIs that cater to strategy, the BARCH model defines the model in relation to the survivability and viability of the business. This is something that really links the business with the model. Sometimes, strategy can be wrong and if a model is built based on that, it will be disastrous. It was at this stage that I began to build into the BARCH model the strong inter-relationship between the factors and how they link up to form innovative approaches and solutions.

5.3 Logistics and Parcel Delivery Company (Choy, Ma and Koo[12], 2012)

The following section is an extract from my research paper on a Logistics and Parcel Delivery Company KPI development work that is published in a conference in 2013. With increasing globalization and the growth of digital media such as Facebook and twitter, companies across industries are confronted with dynamic competition conditions. This change in the external environment changes many business processes which can be negative for some businesses. For

[12]Ma and Koo are both co-authors as they are the reviewers and thus included as part of protocols.

example, in the 2007 Global Financial Crisis, small and medium companies faced credit crunch crisis where they are unable to obtain credit to expand or maintain their business due to a lack of liquid credit. This led to a change in their business model and processes. Such change can have drastic effects on their business and the ongoing changes to the business process can be devastating to the business.

To counteract such dynamic environmental changes, most businesses attempt to model, automate, optimize and monitor their business processes. This has resulted in an increased acceptance and adoption of business process management systems (BPMS). This development has been further encouraged by the increasing number of regulatory requirements imposed on business operations across a wide spectrum of industries as a result of various scandals. The most prominent and powerful legislations passed are the Gramm-Leach-Bliley Act (US, 1999) and the Sarbanes Oxley Act (SOX) (US, 2002). These acts controls and regulate any entities which have operations in the United States of America. While these regulations have coverage across industries, there are also industry-specific regulations such as the Basel III accord (Basel, 2011) or the European Money Laundering Regulation (UK, 2003). Demonstration of compliance with specific legal legislations and international standards often requires the company to document their existing operating business processes as well as verifying that the current set of processes conform to legislation and are strictly followed.

In the case of the logistic company, they have several types of delivery services with numerous business processes controlling the operation of the company. This is a tightly regulated business and they have to deal with authorities such as port customs and immigration customs. Because of the nature of the business, the legislative requirements force the company to formulate their existing business process around these legislations. This renders a great deal of burden on the company. At the same time, the nature of the service rendered requires specific contractual agreement on the service level that makes it important to analyse the business processes for any anomalous behaviours. This is critical as certain dangerous goods can be a threat to the public if they are lost along the way.

During the process of understanding the business, it is obvious that the BARCH model gives a very detailed understanding of the key operations and environment conditions. DELTA framework will analyse the problem from an enterprise level and the processes involved. The most important factors that need to be included are the complexities of the business regulations and how they tie in with the cost structure. The cost and revenue of the business is closely linked to the operation as compared to the previous two cases. The regulation implies that there are financial penalties for infractions and the penalties can easily wipe out the revenue. Thus the regulation is an additional layer of cost on top of the business. It is like a constant bleeding wound in the company that slowly drains it. This reinforces the importance of cost and revenue in the model. In this case, while human still plays a factor, it is not as crucial as the previous ones.

Currently for the daily operation, the business processes are mapped out using the BPMS software and 90% of the bookings for services are done electronically. The operation uses electronic devices to scan the items during the delivery process through wireless data connection that track the movement of items from allocation to receipt. The devices capture all of the process phases and statuses and update them to the system through the wireless connection. This ensures that the data is captured in real time situations. However, there is a lack of visibility to the compliance level of the

business processes. The existing BPMS does not allow the company to check their existing processes. This posed a severe operational risk to the company should there be any violation in the compliance. At the same time, this is the first case that I encountered where the Analytics solution did not address the business needs. The solution only addressed the need to track information but does not report the information.

Most enterprises that operate in heavily regulated industries, such as financial services or health care, are controlled by huge number of regulatory requirements that define most of their operations. As these implemented requirements needed to be enforced by a multitude of internal business and IT controls, several regulations recommend the use of industry standards, such as COBIT (Control Objectives for Information and Related Technologies) (COBIT, 2005) and ITIL (Information Technology Information Library) (ITIL, 2006), in the implementation of any enterprise IT system. These standards are comprised of well-defined abstract process definitions that can be adjusted to match individual needs. Typically standards are useful when they are applied to standardize processes.

Most approaches will implement the deontic logic in the system that checks whether the flows of the business processes are logical or not. While a popular solution due to the ease and speed of implementation, the solution does not take into consideration the entire process chain and only focuses on the immediate events in the process chain. This can be problematic as the business might be stuck in a loop due to logical errors without the business being aware of it. At the same time, it is difficult to compare the entire business process in question that might not be legally compliant but logically feasible in certain steps.

The company in this case will require a solution that combines both logical checks and business process checks. To assist the company to solve the problem, we applied the BARCH model to develop the nomenclature solution. Using the new nomenclature approach, we mapped out the existing processes and converted them into general processes. We have noted that there areseveral possible repeated processes and sub-processes to account for all these, thus we have generated several iterations of the different processes. Using the nomenclature forms, we analyse their existing structures. This is also the first time that the BARCH mode is used to develop innovative solutions rather than analysing problems. Instead of using BARCH on the business problem directly, we actually applied BARCH on the old solution to analyse its flaws to come up with a new solution. This reinforces the concept that Analytics is actually applicable in many areas and it is up to the creativity of the user.

Nomenclature is defined as a systematic approach to the naming of items in the area of science or arts by individuals or community. It can also refer to the systematic naming of items according to taxonomy. In most scientific disciplines, there are established standards of nomenclature. Nomenclatures also allow for easy identification of similar components between two chemicals or species.

In process management, most of the business processes were mapped out in BPMS software in a logical decision flow manner such as the use of BPELs (Business Process Extraction Language) (Antoniou et. al., 2005; Ghose and Koliadis, 2007), BPCL (Business Process Compliance Language) or BPSL (Business Property Specification Language) . These models are extremely useful in understanding whether the flow is sensible and what the components of the process are. They rely

heavily on defining a logical pathway and whether the pathway makes sense. Other software uses symbolic logic such as Standard Deontic Logic (Alberti, 2004; Alberti, 2005) to assist in the checking of the process.

Below we will define the generalized structure for a process.

$$\dots\{a \Rightarrow b\}[n] \dots$$

where a is the start process and b is the end process with n representing the number of times that a process has been repeated sequentially. Thus assuming that a process is repeated for 3 times, below is the form that it will take.

$$\dots\{a \Rightarrow b\}[3] \dots$$

The same can be done for processes with nth components occurring m times.

$$\dots\{a_1 \Rightarrow a_2 \Rightarrow \cdots a_{n-1} \Rightarrow a_n\}[m] \dots$$

While we have discussed the case where processes are repetitive, the case for single event repetition has not been discussed. To enable one to distinguish a single event from repeated process, we will be using the notation as below.

$$\dots a(n) \dots$$

The notation used are simple and comprehended easily. Below is a process flow example from a logistic company

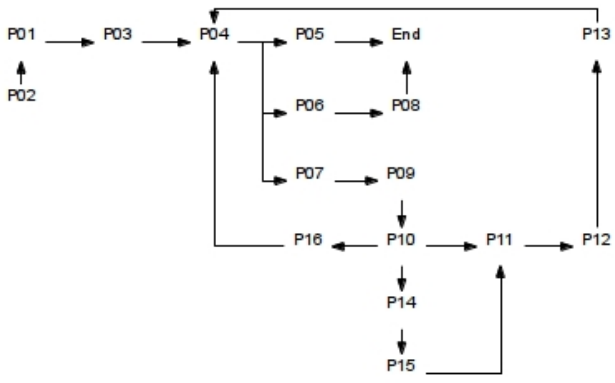


Diagram 1: Process Flow 1 for a logistic company

From the process, we can see that the existing process is extremely complicated and there are many sub-processes and events. The process also indicated major repetitive cycles. However, the diagram does not give the following information:

- Are any of the events self repetitive?
- Are any of the sub-processes self repetitive?

The other problem is that the process cannot be used directly to determine in an operation system whether any of the processes have deviated from the process flow diagram. Thus to solve this issue, we develop an algorithm to simplify the processes into a generalized form. Reader is referred to the paper Choy et. Al. (2014) for further reading.

In the case of the logistic company A, they have several delivery services with many business processes controlling the operation of the company. Because of the nature of the business, certain legislative requirements force the company formulate their existing business process around the legislation. At the same time, because of the nature of the service rendered, there are contractual agreements on the service level making it important to analyse the business processes for any anomalous behaviours.

The key revenue sources of the companies are through the delivery charges and fees of the delivery. Thus the process has to be as efficient as possible to maximize the number of transactions or deliveries as possible. However, the cost of doing this business is quite significant, as it requires an extensive amount of investment in the hardware and vehicles. At the same time, it is important for compliance to be in place to avoid any disputes that might lead to various types of claims against the company. Cost of claims can vary depending on the types of item delivered and whether the compliance process is in place. If there is no proper process, delivery packages might be lost due to human process. This is also the first project where there are strong links between the business and cost. Unlike previous cases where cost is a product of business, cost here is an essential part of business and revenue. The claims are an unknown cost that can occur any time.

Currently, the business process is mapped out using the BPMS software and most of the bookings of services are done electronically. However, there are no visibilities to the compliance level of the business processes. The existing BPMS does not allow the company to check their existing processes. This is an example of a solution that does not solve the business problem. Such solutions typically only focus on specific tasks to solve and does not attempt to address the business needs holistically. This reinforces the idea that a holistic model or framework like BARCH is needed.

Without using the generalized form, there are almost 30,222 different processes that needed to be mapped using the BPMS system. This is tremendous amount of work and passing them through the system individually to test the logic flow is computational intensive work. Such intensive work is also a major source of cost to the organization. Using the generalized nomenclature approach, we reduce the processes to 10,272 that is around 35% of all the processes. This reduces the time needed for verification by 65%.

The nomenclature approach offers improvement to the compliancy checking in the following area.

1. Ease of logical interpretation

2. Detection of non-compliancy

3. Real time detection is possible

The approach allows for flexibility to users who want to modify for their company usage. The ease of applying the compliancy check in operation also enables the management to have good visibility on compliancy as well as access to near real time reports of any compliancy failures.

Learnings

The BARCH model is used to solve the problem. As a company that is controlled by regulations, the unique environment has various effects on the operating cost structure. In this case, the business has been addressed by an Analytics solution that does not meet the company's requirements. The problem presents an interesting application of the BARCH model to a specific problem of an Analytics solution rather than the business. The complexities of the cost structure and how it affects the revenue also reinforces the need to incorporate the two factors in the BARCH model and linking them up with the business. While human factors are present, they are not as pervasive as in the previous two examples. This weakens the human factor in the BARCH model and it is important to consider cases where the business model does not incorporate m many human elements.

Even though human elements are critical to all businesses, they are not the only drive and in fact, factors such as cost and revenue typically hold more sway in management decisions than human factors in times of crisis. There were also an extensive amount of regulatory requirements in the project that presented a new paradigm to the BARCH framework. While it is tempting to put in another factor to account for it, the more appropriate approach was to subordinate it under business. The rationale is that not every industry comes under serious regulatory oversights and in those cases, BARCH can be extended to have a factor incorporated in such cases. The project guided me on the process of handling tightly regulated environments in the context of Analytics and how BARCH can be incorporated natively without modifications.

6. Conclusion

This critical engagement with my papers and my model has been in many ways a challenging one. I have worked with many other experts and Analytics practitioners in the field who reviewed my work in earnest details. My training in maths and economics had pulled me towards a positivistic view of knowledge and my profession requires considerable skills in managing logistics, heuristics, data, statistics, technology, software. However in the process of reviewing in my practice why certain things did not work I found myself developing a model for Business Analytics which came to be more cognisant of the value of the human aspect. This was already acknowledged in the field in that the success of Analytics depends on the quality of the data input. Data about human behaviour cannot be left out. This is a challenge because human behaviour is not homogeneous. Through my projects I became convinced that a model was needed that allowed for a level of adaptability to each unique context of the organisation and the behaviours of the team most responsible for that set of processes or activities.

The reviews of my work and the concept of expertise have also given me new learning and directions in my professional practice. The works done have benefitted the organizations and have led them to develop their business Analytics capabilities. The parcel delivery company that has implemented the process checker has done an organizational restructuring that has put business Analytics as one of the pillars of transformation for the organization. The Basel Committee in their consultation paper (BIS, 2015) has acknowledged similar problems in the credit risk modelling practices with regards to the various definitions, which was addressed by my work. The airport terminal has incorporated business Analytics into their core practices and has redeveloped their operations to further optimize utilization after the work was presented.

After the deep discussion about the concept of expertise, I have developed a strong concept of the meaning of expertise and what it entails to be an expert. This is currently being used in my work with the various user groups to determine the appropriate people to appoint as experts in their specialized domains. The expertise discussion also highlighted the pathway for analysts to progress to experts. The expertise argument also further developed the concept of Episteme, Techne and Phronesis which became the building block for development of analysts. The three concepts also became the criteria to evaluate experts and the level of expertise. Through my works with colleges and tertiary institutions, I am slowly working with them to develop education material and pathways that will present a new learning paradigm and framework for business analysts. Some of my work has received awards [13]and students trained using the methodology has also performed well on the international stage[14].

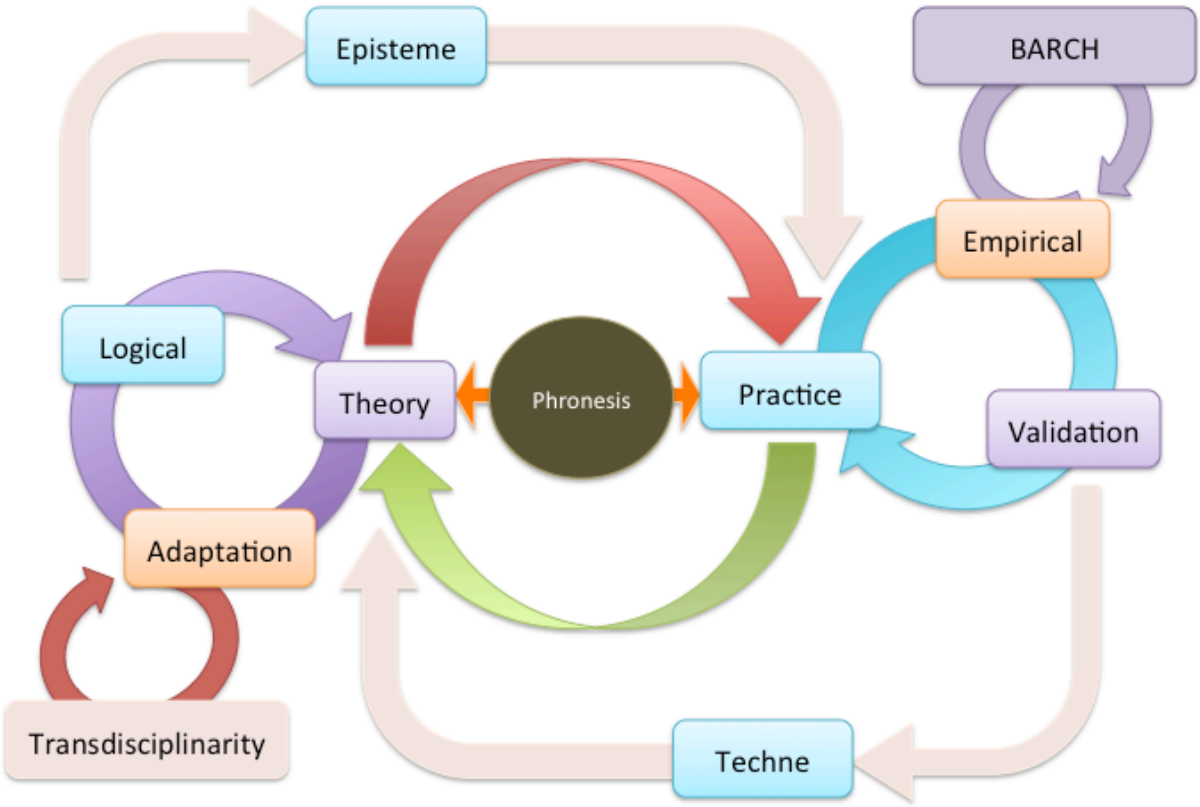
[13]http://cio-asia.com/resource/applications/singapore-academics-bag-teradata-corps-inaugural-tun-award/
[14]https://www.google.com.sg/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&cad=rja&uact=8&ved=0CDYQFjAI&url=http%3A%2F%2Fwww.smu.edu.sg%2Fnews%2F2013%2F04%2F30%2Fsmu-congratulates-cally-ong-being-selected-sas-student-ambassador&ei=i3uxVOaWPNOTuASeh4DgAw&usg=AFQjCNE0ILugturL6KTMVwh2-RQXAzg3_w&sig2=_7LZ-V6lh7XWnalZqnp3zw&bvm=bv.83339334,d.c2E

Carrying out this critique has confirmed for me the value of this human element for the progression of Business Analytics as a field but also for the progression of businesses in an increasingly crowded and competitive world with almost exponential layers of complexity. We cannot approach complexity with an inflexible tool just as we cannot repair all deficiencies in a passenger aircraft with a hammer. During the process, I also recognize the beginning of several major failures in Analytics [15]projects that were once touted as the beacons of analytical successes[16]. One of the key problems mentioned was the behavioural aspects of human beings and how the models did not incorporate those elements and were entirely driven by data collected from the human searches without considering the dynamic relationship between the elements. If BARCH had been applied, the modeller would have noticed the chasm between the business and the human element and that any such model would be misleading once the behaviour changed.

I also became more appreciative of the fact that I could not have developed a new framework without a thorough understanding of the existing models to see what it was they were failing to address. This process of reviewing models and literature is part of the epistemic process that enriches knowledge of the various approaches to solving problems on the one hand and improving my individual practice on the other. This process was also responsible for helping me to conceptualise and articulate this model to my peers and to clients. It has moved my thinking on relating to the evolving nature of Business Analytics. The discipline is truly a transdisciplinary subject at its core.

The discussion about the nature of Analytics also provides me with many interesting insights to the concept of transdisciplinarity. Before I engage in this critical review of the practice of Analytics, I was one of the many Analytics practitioners who are not actively concerned about the discipline and nature of Analytics. However, the examination and discussion is actually critical to the development of the discipline. By determining the nature of the subject, it helps to understand how the discipline should behave and what the key characteristics that define the practice are. The most important learning is the importance of context to transdisciplinary practices, which is incorporated into the BARCH framework in the form of the Business aspect. Other frameworks address the contextual elements as discrete aspects compared to BARCH. The Business Analytics framework incorporates business processes, the GREAT framework uses relevance and timeliness, while the DETLA framework incorporates the Enterprise aspect. These frameworks, in my opinion, lack the comprehensive integration of the context in the formulation process.

[15]http://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu
[16]http://time.com/23782/google-flu-trends-big-data-problems/



BARCH also addresses the lack of boundaries in transdisciplinarity practices. BARCH acknowledges the inherent limitations of positivistic or phenomenological approaches and recommends the use of both in the model. While on the surface, this looks more interdisciplinary or multidisciplinary, the need to incorporate the various complex elements in business makes it impossible for cross applications of disciplines. Any approaches or solutions will need elements from each discipline combined and integrated into one so as to fulfil all the requirements of the problem. Throughout the case papers, we can see that Analytics practices require combining and integrating the theories from various disciplines. The other frameworks, I would argue, do not address this aspect of the practice and simply use the disciplines in a non-integrated manner.

The discussion about BARCH also changed my view of practice and theory in business Analytics. Through the lens of Epsiteme, Techne and Phronesis, I see the complementary nature of practice and theory in business Analytics. Business Analytics does not yet have a core nature or discipline or theory that guides the development of the field. It adapts theories from other fields by logical formulation of theory that gets validated in practice. In practice, we need skills to empirically validate the knowledge that develops practical skills that generate the results that are then adapted back into theory. Through the cycle of theory and practice, phronesis is achieved. No longer bounded by disciplines and the need for a core theory, the review and engagement in the public work has brought about new perspectives on problems that I need to solve and how I and other analysts can adapt theories from other fields to solve problems. The cycle of theory and practice also brings about new questions that we, as analysts, need to answer about the field such as the role of ethics, the value of benefits and the nature of knowledge. Further works in these areas will be valuable to Analytics practitioners to understand their work better and to avoid any pitfalls.

The development of the model required me to increase my technical skills or Techne in the model. My Techne and Episteme in the practice of Analytics will also help me to refine the model to make it consistently relevant to changing contexts and practices. The success of the model has given me the confidence to go on challenging my peers in this field to be more flexible. As an expert, I am on several advisory boards for Analytics practitioners and currently working with these organizations to develop relevant training materials for the Analytics practitioners as the field develops. The problems faced during the various projects also generated in me a level of Phronesis - practical wisdom. I developed a deeper understanding of the practice and how to use my model to solve progressively complex problems. I also share my knowledge, technique and experience with other Analytics practitioners through my association and user groups.

I developed the framework and share it with many analysts as well as managers who are in the process of implementing business Analytics. During my consulting and project work, I have received good feedback from businesses who find my explanations simple and easy to digest. This is because I learn their business and understand what is relevant to them in their context. This understanding allows me to structure the problem in a way that they can understand and appreciate the solution presented. The model allows me to appreciate the knowledge each of the staff can bring to make the Analytics work for them. This appreciation also encourages their support and participation. As an expert, my role is to guide new Analytics practitioners and train them to be better analysts. This is supplemented by an evangelical role for business Analytics in organizations and how to implement business Analytics to solve their problems.

As part of my pro bono work, I am also preparing an e-book that helps analysts to perform their job better. The framework has also helped me to advise start-ups in the area of Analytics which hopefully will spin off into something bigger. The first start-up that I advised on was on a pro-bono basis. I applied my model to solve their start up problems. They won the best financial start-up award. They also went from a start-up to a company with 50 million USD valuation company in 9 months. The second start up that I am advising now and which I am now a part of the senior management, has secured a 20 million investment with a 100 million valuation. The second start up will be using my model as the foundation of their Analytics solutions. The discussions on the concepts of Doxa, Endoxa and Gnosis also helps me to formulate new Analytics products and solutions that the new company will be producing as part of their product line ups.

To summarize, my entire journey in doing this critical review of my work of an applied framework that has been proven in practice is a process that has improved my knowledge, skills and understanding of my practice to a higher level and given me the opportunity to re-examine the discipline at its very core. The journey gave me the opportunity to reflect on the work and how my life experience changed and moulded my outlook as well as the framework. The path to expertise requires hard work and commitment and continuous professional development in how practice informs theory and the reverse. Experiential learning is at the core of the evolving business analyst. Unlike most other disciplines where theoretical work and armchair theorising is common, business Analytics requires fieldwork and direct work experience to develop expertise. In the days of Linkedin, everyone is a self-proclaimed expert. My reflections on expertise also enabled me to develop a framework to assess the expertise level of Analytics practitioners that is important in distinguishing their capabilities for the organization. The work on transdisciplinarity has given me a conceptualisation that helps me to articulate the complexity of not only what my model does but

what Business Analytics does, an evolving and flexible framework in a world of increasing complexity in which all disciplines involved are changed by the encounters. This gives me confidence to go on involving more disciplines and to help many others in their work and to contribute to preparing business analysts of the future, for the future. I would like to quote the closing paragraph of *Transdisciplinarity as Translation,* a chapter in Transdicsiplinary Professional Learning and Practice (Maguire, 2015).

Transdisciplinarity in professional studies doctorates aims to go beyond the 'strait-jacket' of mere problem solving into an era that does not negate disciplines and dilute them into some kind of epistemological soup but rather creates the conditions for more metanoic solutions to managing complexity and the liberating of thinking and action from hegemonic island paradigms. These may be disciplined bound in higher education but in the world of markets, resources and political manoeuvring in which profit and power are synonymous, the hegemonic islands are global companies and super-institutions with vested interests and therefore have more power to exclude, marginalise or reduce the share in the future of large sections of the inhabitants of the planet. A transdisciplinary approach can do in the new hegemonic islands what it has started to do in research education and practice.(Maguire in Gibbs 2015:176)

References

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, *32*(8), 10-14.

Balsiger, P. W. (2004). Supradisciplinary research practices: history, objectives and rationale. *Futures*, *36*(4), 407-421.

Baltes, P. B., Gluck, J., & Kunzmann, U. (2001). Its Structure and Function in Regulating Successful Life Span Development. *Handbook of positive psychology*, 327.

Baltes, P. B., & Kunzmann, U. (2003). Wisdom. *Psychologist*, *16*(3), 131-133.

Baden-Fuller, Charles; Mary S. Morgan (2010). "Business Models as Models". Long Rang Planning 43 (2/3): 156–171.

Baden-Fuller, C., Demil, B., Lecoq, X., & MacMillan, I. (2010). SPECIAL ISSUE Business Models.

BIS, 2015. *Developments in credit risk management across sectors: current practices and recommendations, Consultative Paper.*

Bose, R. (2009). Advanced Analytics: opportunities and challenges. *Industrial Management & Data Systems*, *109*(2), 155-172.

Berger, M. T. (1997). The triumph of the East? The East Asian miracle and post-cold war capitalism. *The rise of East Asia: Critical visions of the Pacific century*, 260-87.

Casadesus-Masanell, R., & Ricart, J. E. (2010). From strategy to business models and onto tactics. Long range planning, 43(2), 195-215.

Chen, Hsinchun, Roger HL Chiang, and Veda C. Storey. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS quarterly* 36.4 (2012): 1165-1188.

Chiang, R. H., Goes, P., & Stohr, E. A. (2012). Business intelligence and Analytics education, and program development: a unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems (TMIS)*, *3*(3), 12.

Choy, M., & Laik, M. N. (2011). A Markov Chain approach to determine the optimal performance period and bad definition for credit scorecard. *Research Journal of Social Science and Management*, *1*(6).

Murphy Choy, M. N. L., & Shung, K. P. Real Time Process Compliance Checking using Nomenclature Approach.

Collins, H. M., & Evans, R. (2002). The third wave of science studies studies of expertise and experience. *Social studies of science*, *32*(2), 235-296.

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The new science of winning*. Harvard Business Press.

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*. Harvard Business Press.

Davenport, T. H. (2010). Business intelligence and organizational decisions.*International Journal of Business Intelligence Research (IJBIR)*, *1*(1), 1-12.

Dictionary, O. E. (2013). Oxford English Dictionary.

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision expertise in real world contexts. *Organization studies*, *26*(5), 779-792.

Dunne, J. (1993). *Back to the Rough Ground: Phronesis and'Techne'in Modern Philosophy and in Aristotle* (Vol. 11). University of Notre Dame Press.

Eikeland, O. (2008). *The ways of Aristotle: Aristotelian phronesis, Aristotelian philosophy of dialogue, and action research* (Vol. 5). Peter Lang.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.

Fine, G. (2003). Plato on knowledge and forms: selected essays.

Frank, A. W. (2012). Reflective Healthcare Practice. In *Phronesis as Professional Knowledge* (pp. 53-60). SensePublishers.

Franks, B. (2012). *Taming the big data tidal wave: Finding Opportunities in Huge data streams with advanced Analytics* (Vol. 56). John Wiley & Sons.

Gary, J. E. (2008). The future according to Jesus: A Galilean model of foresight.*Futures*, *40*(7), 630-642.

George, G., & Bock, A. J. (2011). The business model in practice and its implications for entrepreneurship research. *Entrepreneurship theory and practice*, *35*(1), 83-111.

George, G., & Bock, A. J. (2012). *Models of opportunity: How entrepreneurs design firms to achieve the unexpected*. Cambridge University Press.

Gettier, Edmund L. "Is justified true belief knowledge?." analysis (1963): 121-123.

Gibbons, M., & Nowotny, H. (2001). The potential of transdisciplinarity. In*Transdisciplinarity: joint problem solving among science, technology, and society* (pp. 67-80).

Greeno, J. G. (1989). On The Nature of Competence: Principles for Understanding in a Domain©. *Knowing, learning, and instruction: Essays in honor of Robert Glaser*, 125.

Gries, P. H., & Peng, K. (2002). Culture clash? Apologies east and west.*Journal of Contemporary China*, *11*(30), 173-178

Gu, J., Gao, R., Li, L., Zhu, Z., & Song, W. (2010, August). Knowledge Inheritance in Traditional Chinese Medicine (TCM). In *Proceedings of the 54th Annual Meeting of the ISSS-2010, Waterloo, Canada* (Vol. 54, No. 1).

Hammer, M., & Söderqvist, T. (2001). Enhancing transdisciplinary dialogue in curricula development. *Ecological Economics*, *38*(1), 1-5.

Hedman, J., & Kalling, T. (2003). The business model concept: theoretical underpinnings and empirical illustrations. *European Journal of Information Systems*, *12*(1), 49-59.

Hibbert, K. (2012). Cultivating Capacity. In *Phronesis as Professional Knowledge* (pp. 61-71). SensePublishers.

Higgs, J. (2012). Realising practical wisdom from the pursuit of wise practice.*Phronesis as professional knowledge: Practical wisdom in the professions*, 73-85.

Horlick-Jones, T., & Sime, J. (2004). Living on the border: knowledge, risk and transdisciplinarity. *Futures*, *36*(4), 441-456.

Hofstede, G. (1993). Cultural constraints in management theories. *The Academy of Management*

Executive, *7*(1), 81-94.

Iriye, A. (2002). *Global community: The role of international organizations in the making of the contemporary world*. Univ of California Press.

Jessica Casey (2013). Seattle's Predictive Policing Program. http://datasmart.ash.harvard.edu/news/article/using-predictive-policing-to-reduce-crime-rate-189

Jetton, T. L., & Alexander, P. A. (1997). Instructional importance: What teachers value and what students learn. *Reading Research Quarterly*, *32*(3), 290-308.

Johnson, M. W., Christensen, C. M., & Kagermann, H. (2008). Reinventing your business model. Harvard business review, 86(12), 57-68.

Jones, B. (1975). An Introduction to the First Five Chapters of Aristotle's" Categories". *Phronesis*, 146-172.

Juengst, E. T. (1995). The ethics of prediction: Genetic risk and the physician–patient relationship. *Genome science and technology*, *1*(1), 21-NP.

Maguire, K. (2015). Transdisciplinarity as Translation. In *Transdisciplinary Professional Learning and Practice* (pp. 165-177). Springer International Publishing.

KEMMIS, S., & SMITH, T. J. (2008). 2. PERSONAL PRAXIS. *Enabling praxis: challenges for education*, *1*, 15.

Kenny, R., Pierce, J., & Pye, G. (2012, January). Ethical considerations and guidelines in web Analytics and digital marketing: a retail case study. In *AiCE 2012: Proceedings of the 6th Australian Institute of Computer Ethics conference 2012* (pp. 5-12). Australian Institute of Computer Ethics.

Kinsella, E. A., & Pitman, A. (2012). *Phronesis as professional knowledge: Practical wisdom in the professions* (Vol. 1). Springer.

Ko, Ryan KL, Stephen SG Lee, and Eng Wah Lee. "Business process management (BPM) standards: a survey." *Business Process Management Journal* 15.5 (2009): 744-791.

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging trends in business Analytics. *Communications of the ACM*, *45*(8), 45-48.

Kohli, Rajiv, and Varun Grover. "Business value of IT: an essay on expanding research directions to keep up with the times." *Journal of the association for information systems* 9.1 (2008): 1.

Krioukov, A., Goebel, C., Alspaugh, S., Chen, Y., Culler, D. E., & Katz, R. H. (2011). Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities. *IEEE Data Eng. Bull.*, *34*(1), 3-11.

Krumeich, J., Burkhart, T., Werth, D., & Loos, P. (2012). Towards a Component-based Description of Business Models: A State-of-the-Art Analysis.

Langford, J. M. (1999). Medical mimesis: healing signs of a cosmopolitan" quack". *American Ethnologist*, *26*(1), 24-46..

Laursen, G. H., & Thorlund, J. (2010). *Business Analytics for managers: Taking business intelligence beyond reporting* (Vol. 40). John Wiley & Sons.

Ma, N. L., Choy, M., & Cheong, M. (2012, July). Uncovering interesting business insights through the use of data Analytics in Airport operation: an empirical study. In *SRII Global Conference (SRII), 2012 Annual* (pp. 803-810). IEEE.

Macklin, R., & Whiteford, G. (2012). Phronesis, aporia, and qualitative research. In *Phronesis as Professional Knowledge* (pp. 87-100). SensePublishers.

Maguire, K (2015). Transdisciplinarity as Translation. In *Transdisciplinary Professional Learning and Practice* (pp 165-176). Springer.

Melrose, S. (2005). Words fail me: dancing with the other's familiar.

Melrose, S. (2011). A cautionary note or two, amid the pleasures and pains of participation in performance-making as research.

Mestrovic, S. (2004). *The Balkanization of the West: The Confluence of Postmodernism and Postcommunism*. Routledge.

Merton, R. K. (Ed.). (1976). *Sociological ambivalence and other essays*. Simon and Schuster.

Naranjo, C. (1972). *The one quest*. Viking Press.

Nicolescu, B. (2005). Towards transdisciplinary education. *TD: The Journal for Transdisciplinary Research in Southern Africa*, *1*(1), 5-15.

Nudurupati, Sai S., et al. "State of the art literature review on performance measurement." *Computers & Industrial Engineering* 60.2 (2011): 279-290.

Osterwalder, A. (2004). The business model ontology: A proposition in a design science approach.

Osterwalder, A., Pigneur, Y., & Tucci, C. L. (2005). Clarifying business models: Origins, present, and future of the concept. *Communications of the association for Information Systems*, *16*(1), 1.

Ramadier, T. (2004). Transdisciplinarity and its challenges: the case of urban studies. *Futures*, *36*(4), 423-439.

Schmitt, C. (1996). The Concept of the Political. 1932. *Trans. George Schwab. Chicago: U of Chicago P*.

Smith, Richard (2014). Truth in the face of change. http://richardsmithpov.com/truth-in-the-face-of-change/

Schön, D. A. (1987). Educating the reflective practitioner: Toward a new design for teaching and learning in the professions. *San Francisco*.

Sharma, R., Reynolds, P., Scheepers, R., Seddon, P. B., & Shanks, G. G. (2010, January). Business Analytics and Competitive Advantage: A Review and a Research Agenda. In *DSS* (pp. 187-198).

Shuen, Amy. *Web 2.0: A Strategy Guide: Business thinking and strategies behind successful Web 2.0 implementations*. " O'Reilly Media, Inc.", 2008.

Slade, Sharon and Prinsloo, Paul (2013). Learning Analytics: ethical issues and dilemmas. *American Behavioral Scientist,* 57(10) pp. 1509–1528.

Sternberg, R. J. (1998). A balance theory of wisdom. *Review of general psychology*, *2*(4), 347.

Stokols, D. (2006). Toward a science of transdisciplinary action research.*American journal of community psychology*, *38*(1-2), 63-77.

Stubbs, E. (2013). The Value of Business Analytics. *Business Analytics: An Introduction*, 1.

Timothy McGrew (2007), Internalism and Externalism, Abingdon, Oxon: Routledge, chapter 1

Teece, D. J. (2010). Business models, business strategy and innovation. *Long range planning*, *43*(2), 172-194.

Turner, S. (2001). What is the Problem with Experts?. *Social studies of science*, *31*(1), 123-149.

Rousseau, Denise M., Joshua Manning, and David Denyer. "11 Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses." *The academy of management annals* 2.1 (2008): 475-515.

Appendices

Appendix A

7-2013

Performance Measurement Design for a Parcel Delivery Company

Junyu Choy
Singapore Management University, murphychoy@smu.edu.sg

Nang Laik MA
Singapore Management University, nlma@smu.edu.sg

Ping Shung Koo
Singapore Management University, pskoo@smu.edu.sg

Performance Measurement Design for a Parcel Delivery Company

Murphy Choy, Ma Nang Laik and Koo Ping Shung

***Abstract*—There were extensive researches on the topic of performance management in various organizations across multiple fields. Literature on performance measurements in logistics can be divided into specific measures and their application in the context or complete framework for performance measurements. In this paper, the focus of the discussion will be the formulation of the framework which handles performance measurements for package delivery service and how the metrics measure the performance and their application in the context of package delivery service.**

***Index Terms*—Analytics, SERVQUAL, Performance Management, Parcel, Logistics**

I. Introduction

Package delivery services play an important role in providing communications and transfer of items in an economy. The traditional postal service of physical delivery of communication items such as mail has been superseded by package delivery, courier services and other auxiliary services. The privatization movement has changed the postal service where public postal services are deregulated and forced to compete with one another. The decline of physical post service and deregulation has forced the postal service to evolve competitively with strong focus on the market and customer needs. While pricing remains a powerful tool in maintaining competitive edge, package delivery services requires other qualitative factors to achieve the edge [1][9].

Due to the market potential of the package delivery service, proper performance measurement is needed in order to improve the overall service level. Many postal companies around the world have been implementing new approaches to support innovative operational practices that maintain or improve their market share [4]. The parcel delivery service consists of carriers that transport items that can be handled by one person [14]. In the context of logistics, the parcel delivery service is commonly considered as part of third-party service provider that ensures a smooth movement of goods within the supply chain [19]. Benchmarking techniques such as analytic hierarchy process are popular as they are able to incorporate both qualitative and quantitative measures. However, the qualitative aspects of the methodology require a well thought framework in order to ensure consistency. While benchmarking approaches appear to be quite popular, the performance measurement approach remains widely popular among the practitioners. This approach is well suited to the postal delivery service which has strong service level agreement that can be used to drive service performance. The choice of service providers is strongly dependent on the service qualities provided by the service providers [3][19]. The key factors include the ability to maintain and maximize level of service, increasing the service coverage provided, and niche market specialization.

There were research that demonstrated service quality improvement is a must [6] for any providers to attain competitive advantage and failure to do so would lead to competitive disadvantages [8]. Thus, any approaches towards effective service quality will correlate with good performance for any service industries and are measured to gain customer satisfaction [15]. In this paper, we will be adapting the SERVQUAL model to measure customer satisfaction towards service quality and relate them to performance measurements [12][15].

II. SERVQUAL Model and Customer Satisfaction

Services are intangible [2] due to them being performances instead of physical objects. Precise specifications for the performances do not work in the same way as the specifications set in physical goods. Services cannot be counted, measured directly or tested ahead for quality assurance. Unlike physical goods which operate independent of the environment, performance of services can be subjected to environment changes which necessitate adaptations to deliver the service. The intangibility of services makes it difficult for service providers to evaluate their service quality and how well they performed [20].

Services are also highly dependent on the delivery of the service. As services are performances, they are ultimately dependent on the service provider which involves human labour. The heterogeneous nature of human labour as well as human interactions makes the the consistency of the delivery of the services difficult to measure. The human factor also creates a layer of uncertainty between the intended service delivery and the actual service delivered.

Lastly, services are produced and consumed simultaneously [7][18]. Due to this simultaneous nature, quality in services cannot be created at production and then delivered to the consumer separately. The delivery requires human interactions which is very difficult to control [11]. Even if the delivery of the service is well controlled due to

Manuscript received Feb 22, 2013; revised Mar 30, 2013.

Murphy Choy. Author is with the School of Information System, Singapore Management University (e-mail: murphychoy@smu.edu.sg).

Koo Ping Shung. Author is with the School of Information System, Singapore Management University.

Ma Nang Laik Author is with the School of Information System, Singapore Management University

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

excellent delivery training, the consumers' participation in the delivery process can affect the final service delivery quality. The consumer's aspect becomes even more important in the cases where consumer's participation is needed to complete the delivery.

In the literature discussions [10][11], there are several main themes:

1. Service quality evaluations are complex and difficult for the consumer than goods quality evaluation.
2. Service quality evaluations are the result of comparing consumer expectations with actual service performance.
3. Service quality evaluations involve both the process of delivery and outcome of a service.

Given the lack of physical and tangible aspects to evaluate service quality [13][16][20], the usual tangible aspects are limited to physical facilities, equipments and personnel. While price is commonly considered to be the pivotal quality indicator in cases where other information is not available, it is not considered to be the main quality indicator or performance indicator. Because of the lack of tangible aspects to evaluate quality, the measurement of quality is therefore tenuous. Some researchers have attempted to measure the gap between expectations and performance as a way to establish service quality. Others measure the quality by evaluating whether the performance has met the expectations [5][17].

The SERVQUAL model is developed [15] to address some of the gaps in the research. The model has 10 major characteristics.

Reliability is defined as the measure of consistency in terms of performance and as well as the dependability. This measure has several interpretations. It could be defined as the firm performing the service right without repeats or the firm honoring its promises. The measures most commonly related with this characteristic are accuracy and timeliness.

Responsiveness is defined as the willingness of the labour to provide services. This measure is strongly related to the timeliness concept used in the previous characteristic with some minor changes. Instead of timeliness, the concept should be the speed of reaction to external stimuli. Common measures for this characteristic are turnaround time and reaction time.

Competence is referred to the acquisition and retention of the skills and knowledge necessary to perform the service. This measure comprises of two components, the first component is the knowledge and the second component is skill. In order to perform any service, the labour must be equipped with the right knowledge. The right knowledge does not guarantee smooth delivery and good delivery requires skill.

Access is referred to ease of contact. This is again another measure of timeliness. Unlike the previous measure, this measure of timeliness refers to the amount of waiting time. If you make any customer wait too long, they will go away.

Courtesy is referred to as the appropriate protocol for customer engagement. The measure may include items such as politeness or friendliness. Any rude or unruly service provider will naturally incur the wrath of customers. However, cross-cultural issues and conflicts may be referred to as a courtesy issue. This is common when the parties have different cultural practices. Courtesy can also refer to image of the company or the labour.

Communication refers to the continuous engagement of the customers in the preferred language. The company has to make sure the services are well explained and that the terms and conditions are understood by the customers. The measure involves the number of language mismatch. The secondary measure is the ease of understanding. Even with the right language, we cannot ensure good understanding of the service and we can measure the number of incidents involving miscommunications.

Credibility refers to the trustworthiness and honesty of the organization. The measure is usually some factors involving the company name or reputation. This is extremely difficult to measure and the most common measure is the number of complaints received. With the advent of social media, the channel has contributed greatly to the measurement of credibility.

Security is defined as the risk – free level. This measure is defined by the probability of loss or stolen items. This can be tracked using lost items. With new data protection acts, the risk has been extended to data privacy issues which need proper data security framework.

Know your customer is an important aspect of business and it is usually defined as understanding the customer's need. This is commonly measured by the amount of information that the organization keeps about the customer as well as analysis done on the behaviour of the customer. Alternatively, it can be measured by how often an update is done on a customer record.

While services do not usually have tangibles, occasionally tangibles such as facilities for the service, receipts and other physical items are considered tangibles. These are measured by customers' reception of them.

The paper is structured with the following sections. In the next section, we attempt to link these measurements of quality with performance measures. In section 4, we discuss about the measures and how they can be implemented. Section 5 discusses the case study and section 6 presents our conclusions.

III. SERVQUAL MODEL AND PERFORMANCE MEASURES

In the previous section, we have discussed about the definitions of quality and how performance is related to quality. Quality in services is determined by the gap between performance and expectations. The SERVQUAL

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

model has defined 10 characteristics that are related to quality and we will be transforming these characteristics to drive performance in package delivery services.

Reliability in package delivery service can be defined as timeliness of the delivery. In most package delivery service, there is a concept of time to deliver which is a form of service level agreement with the customer. Reliability can also refer to the confidence in the delivery before deadline. Reliability can also refer to the confidence in the final delivery. One common measure of reliability of package delivery is the measurement of how many packages that reached the final destinations. Another alternative measure is the number of packages who were sent to wrong destinations. Both measures effectively focused on the final outcome of the delivery and whether the delivery is correct or wrong. Thus reliability embodies many aspects of the performance and quality.

Responsiveness for package delivery service can be defined as speed that the labour can pick up and deliver the package. As mentioned in the earlier section, the measure has relations with the concept of timeliness and more of a reaction time to external stimuli. In this case, the measure would be the measurement between the time of contact to time of package collection or from point of collection to point of delivery. Both measure important information about the performances. The first measure will determine the reaction time to any package delivery request which can be perceived by the customer as the eagerness to engage them as well as their importance. The second measure is basically another form of measure for reliability but can be perceived as the importance of the package to the delivery provider.

The two separate and distinct components of competence require more thorough formulation of measures. The knowledge measure is something that is both internal and external. The internal aspect of the knowledge involves the understanding of the knowledge and how it can be used. This aspect can be tested using tests during training sessions and the scores can then be used as a measure. The external aspect of knowledge can be determined by the capability of the labour to answer any questions directed professionally. This can be measured as an external measure by deriving the number of complaints against any labour for incorrect or insufficient explanations. The skill component can be tested by the number of complaints received for any labour with regards to the management of the delivery process.

Access is referred to ease of contact with the labour and service provider. This measure of timeliness refers to the amount of waiting time for the labour or service provider to react. Too much waiting or extended period of lack of feedback can result in severe frustration for the customers.

Courtesy refers to the appropriate protocol for customer engagement which is usually a subsection of knowledge. The measures include items such as politeness or friendliness towards the customer. Any rude or unruly labour will naturally incur the wrath of customers that will result in complaints and a direct measurement of the etiquette of the labour. The complaints can be through various channels such as phone calls or even complaints on twitter and social media.

Communication in this context refers to the continuous updating of the package delivery status to the customer using the proper medium and language. This measure can be evaluated through the number of complaints received. Constant feedback is also critical and this can be measured by the frequency of the updates to the customer which acts as reassurance of the delivery situation. Communication measures can also be defined as the number of customer call in inquires for a single package.

Security in the context of package delivery service refers to the security of the package. This measure can be defined by the extent of damage to the package. The customers would naturally assume that the packages are well protected and handled gently. Any packages experiencing serious damages would be considered less than acceptable and that no proper security was in place to protect it. The even more serious problem would be the case where the object is opened during the delivery which would imply a breach of security. This kind of incident will require proper investigation. Another possible security issue is the likelihood of improper delivery which resulted in lost packages. These cases are mainly due to labour committing fraud or delivering the goods in an inappropriate manner. Both cases can be measured using complaints.

Credibility, Know your customer and tangibles are aspects of the quality which are not directly applicable in the package delivery service. Any service provider without credibility will not even receive the basic license to provide the service. Package delivery services usually handles customers who are not entirely regular and on ad-hoc basis making it difficult to produce any interesting analysis of the customer's behaviour. As package delivery has no tangibles, there are no tangibles for comparison.

IV. Performance Measures and Implementation

The most useful performance measures are only as powerful as the data driving it. Any proper performance measure implementation requires careful evaluations of the data quality as well as the data content. Improper use of the data to construct the performance measures would result in biases and inaccuracy which might affect the true view of the performances. To ensure good understanding of the measures, we will be defining the various measures and the factors which they correspond to in SERVQUAL framework.

$$\text{Let } i \text{ be the package } i \text{ for delivery}, \forall\, i \in \{1,\ldots,I\}$$
$$\text{Let } d_i \text{ be the intended destination for package } i$$
$$\text{Let } e_i \text{ be the actual destination for package } i$$
$$\text{Let } s_i \text{ be the starting location for package } i$$
$$\text{Let } t_{s_i} \text{ be the pick up time for package } i \text{ at starting location } s_i$$
$$\text{Let } t_{d_i} \text{ be the delivery time for package } i \text{ at destination } d_i$$

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Let t_{c_i} be the contact time for package i

Let t_i be the expected delivery time as defined by the service level agreement for package i

Let k_i be the expected pick up time as defined by the service level agreement for package i

Let x_i be the binary variable for delivery within the agreed service level time as defined below

$$x_i = \begin{cases} 1, & \forall\, i, (t_{d_i} - t_{s_i} \le t_i) \\ 0, & \text{otherwise} \end{cases}$$

Let y_i be the binary variable for correct delivery final destination ,

$$y_i = \begin{cases} 1, & \forall\, i, (e_i = d_i) \\ 0, & \text{otherwise} \end{cases}$$

Let z_i be the binary variable for pick up within the agreed service level time as defined below

$$z_i = \begin{cases} 1, & \forall\, i, (t_{s_i} - t_{c_i} \le k_i) \\ 0, & \text{otherwise} \end{cases}$$

Reliability in package delivery service has been defined using several measures. The first measure refers confidence in the delivery before deadline. This measure can be construed as the proportion of deliveries delivered before the final service level agreement deadline.

$$Reliability = \frac{\sum_{i=1}^{I} x_i}{I}$$

Reliability can also refer to number of packages that reached the final destination d.

$$Reliability = \frac{\sum_{i=1}^{I} y_i}{I}$$

The alternative measure is the number of packages who were sent to wrong destinations. This alternative measure can be calculated as the number of packages that went to a wrong destination or the maximum number of wrong destinations for any package. The first measure is the complement of the previous measure.

$$Reliability = 1 - \frac{\sum_{i=1}^{I} y_i}{I}$$

The second measure is as below.

$$Reliability = MAX\,(I - \sum_{i=1}^{I} y_i)$$

Both measures are effective in measuring the final outcome of the delivery although the second measure in this case is not normalized and cannot be compared easily. Good performances in these measures will indicate excellent service quality as well as performances.

Responsiveness has been defined as speed that the labour can pick up and deliver the package. The more appropriate measure would be the measurement between the time of contact to time of package collection. The measure can be perceived by the customer as a form of engagement. The measure is defined as follow

$$Responsiveness = \frac{\sum_{i=1}^{I} t_{c_i} - t_{s_i}}{I}$$

Measuring competency is very difficult for both the internal and external competency. As mentioned in the previous section, we discussed about the use of tests to validate the internal aspects of the knowledge. While this is easily achievable, it is unclear how the test results will correlate to knowledge. The more appropriate and long term measure is a mixture of both test results and training period provided. The first measure of exam results can be separated into short term, midterm and long term test period. Given that it is period based testing, the measure can be weighted according to the period in the prior testing. The most common weighted moving average is the exponential weighted moving average. Thus we can apply the concept to the measure.

Let u be the time period , $u \in \{1, \ldots, U\}$
where 1 is the latest period , U is the oldest period
Let g_u be the score for time period u
Let a be the weight

$$Competency = a\left(g_1 + (1-a)^1 g_2 + \ldots + (1-a)^{(u-1)} g_u\right) + (1-a)^{(u)} g_u$$

Competency can be calculated as the average number of training days in the past *u* time period.

Let m_u be the number of training days for period u ,

$$Competency = \frac{\sum_{u=1}^{U} m_u}{U}$$

External measures are harder to measure due to the inability to extract the required information easily. To measure knowledge in the external aspect, we can track the number of cases of complaints about the labour's lack of knowledge. A simple and effective measure of competency will be to calculate the number of complaints in the last u time periods.

Let r_u be the number of complaints received for period u ,

$$Competency = \frac{\sum_{u=1}^{U} r_u}{U}$$

The measure can be broken down into complaints about the lack of product knowledge and situations which can be attributed to lack of skills. The latter can then be used to measure the skill level of the labour. All the remaining measures such as Access, Courtesy and Communication can be measured in the same manner.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Security can be measured in terms of the proportion of damaged goods or the amount of compensation for damaged goods.

$$Let\ f_i\ be\ the\ binary\ variable\ for\ damaged,$$

$$f_i = \begin{cases} 1, & package\ i\ is\ damaged \\ 0, & otherwise \end{cases}$$

$$Security = \frac{\sum_{i=1}^{I} f_i}{I}$$

For the case of compensation, we can calculate the average compensation for the last u time periods.

$$Let\ w_u\ be\ the\ compensation\ amount\ for\ period\ u,$$

$$Average\ Compensation = \frac{\sum_{u=1}^{U} w_u}{U}$$

In the next section, we will discuss about some of the measures described here and how they are implemented in the next section.

V. Case Study: Service Delivery Compliance Performance Report

The service package provider needs a performance report on their package delivery service. The provider believes that the existing business process is inefficient and merit further investigation. To facilitate their investigation and improve the performance, they need a report which incorporates all the quality measures that are relevant. However, the database has limited information and certain aspects of quality service such as customer complaints are not recorded. The level of compliance with the existing process is also a mystery to the management and they wish to motivate the labour to comply with the business process.

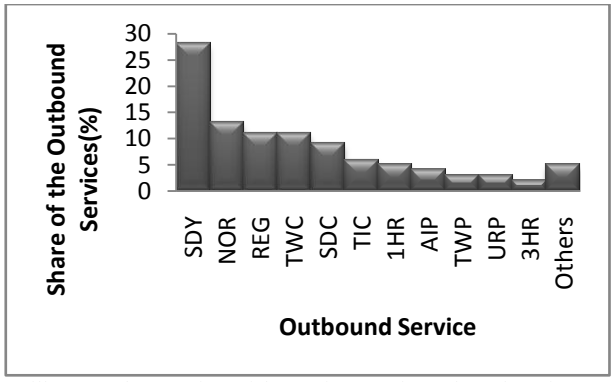


Fig. 1 illustrates the overview of the service type for outbound service – to deliver the item out of the country is shown above.

Using the SERVQUAL framework and the relevant measures, we will attempt to develop a performance report that incorporates the information. The information provided includes the time stamp for each stage of delivery for any packages. Thus we have the time information. The data base also provides the ending process which indicates whether the item has been delivered to the right destination or there was a delivery failure. These two information can be used to generate measures that relate to reliability, security as well as responsiveness.

To measure reliability, we have to consider the appropriate measures to be used. Given that some of the package services have time limits, we can use the measure that calculates the number of deliveries on time. At the same time, since we know the starting time and ending time for each delivery, we can also estimate the maximum and minimum time for the particular business process. In terms of security, we can observe the start and end status as well as validating the existing process according to the business process map for the operation. Using the measures, we came up with the following measures.

1. # Occurrence
2. Start Status
3. End Status
4. Mean time taken
5. Min time taken
6. Max time taken
7. Is the process correct?
8. % SLA satisfied

These measures however, only measure the performance of the labour at the process level and do not represent the entire business operations. However, they do provide insights into business processes which are taking too long, lapse in process compliance or have poor service level agreement fulfilment levels. At the aggregated level for the different business processes, we have the following information as the measures.

1. Type of Service
2. No of Items
3. No of failed delivery
4. No of Delivered Items
5. No of Items considered delivered
6. Average length of process
7. Average time in Hours
8. Maximum time in Hours
9. Minimum time in Hours

At the service level, the measures of failed deliveries and length of process add additional dimensions giving us additional insights to the reliability as well as responsiveness of the various services. Applying the measures to the company, we modified the report to reflect the data that is available for use. Using the information, we develop the following report.

Service Type	Process Flow	# occurrence	%	Start Status	End Status	Mean time taken (Hr)	Min Time taken (Hr)	Is the process correct	% SLA satisfied?
SDY	AC=>TI=>AL=>FD	23729	15%	AC	FD	8.11	4.5	Y	90%
SDY	AC=>TI=>AL=>FD=>BA	20807	13%	AC	BA	5.31	3.1	N	91%
SDY	AC=>AL=>FD	8483	5%	AC	FD	5.14	1.8	Y	98%
SDY	AC=>TI=>AL=>BA=>FD	7553	5%	AC	FD	12.14	2.1	Y	78%
SDY	AC=>AL=>FD=>BA	4820	3%	AC	BA	13.33	1.7	N	71%

Table. 1 illustrates the overview information of the service type SDY for outbound service

Service Type	Process Flow	# occurrence	%	Start Status	End Status	Mean time taken (Hr)	Min Time taken (Hr)	Is the process correct	% SLA satisfied?

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

1 HR	AC=>AL=>FD	10186	36.3%	AC	FD	0.61	0.21	Y	99%
1 HR	AC=>AL=>FD=>BA	9569	34.1%	AC	BA	0.63	0.22	N	98%
1 HR	AL=>FD	2952	10.5%	AL	FD	0.63	0.25	Y	98%
1 HR	AL=>FD=>BA	2903	10.3%	AL	BA	0.72	0.30	N	99%

Table. 2 illustrates the overview information of the service type 1HR for outbound service

Service Type	Process Flow	# occurrence	%	Start Status	End Status	Mean time taken (Hr)	Min Time taken (Hr)	Is the process correct	% SLA satisfied?
3 HR	AC=>AL=>FD	4104	35%	AC	FD	2.531	1.12	Y	95%
3 HR	AC=>AL=>FD=>BA	3781	33%	AC	BA	3.566	1.85	N	81%
3 HR	AL=>FD	1249	11%	AL	FD	2.781	1.91	Y	98%
3 HR	AL=>FD=>BA	1229	11%	AL	BA	2.113	1.31	N	97%

Table. 3 illustrates the overview information of the service type 3HR for outbound service

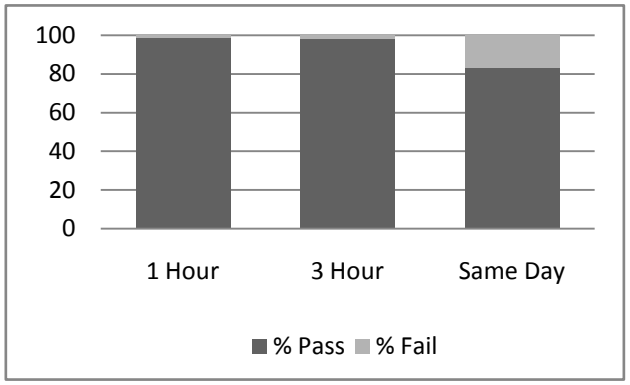


Fig. 2 illustrates the overview failure rate of the service type for outbound service

The report gave much insight to the operation of the company. Among all the services, we have chosen SDY, 1 HR and 3 HR service types which have contributed to about 40% of the total service types and show the detail of the performance report. The first few findings include the appearance of process flows which do not conform to the existing processes. The report also gave insights to the percentage of deliveries made within the SLA limits. Additional insights were derived from the delivery time as well as the number of times an illegal process was found. These information gave insights on the sources of problem and how much impact they are having on the operation.

VI. Conclusion

The SERVQUAL framework can be modified to formulate performance measures which can be used to drive performance. In this paper, we have demonstrated the various facets to performance measures and how they are related to quality measures. The research has also narrowed the scope of the various measures to the key performance measures as well as formulating the various formulas and calculations needed to arrive at the measures. The paper also demonstrated in a small case study on how the various measures can be applied both at a business process level and an aggregated service level. Hopefully, this paper will prompt more research in the relations between quality and performance in package delivery services.

References

[1] Bard, J.F., Binici, C. and deSilva, A.H. (2003), “Staff scheduling at the United States postal service”, Computers & Operations Research, Vol. 30 No. 5, pp. 745-71.

[2] Berry, LeonardL. (1980), "Services Marketing Is Different," Business, 30 (May-June), 24-28.

[3] Bourlakis, M. & Melewar, T.C. (2011). Marketing perspectives of logistics service providers: Present and future research directions. European Journal of Marketing, 45(3), 300 – 310.

[4] Borenstein, D., Becker, J.L. and Prado, V.J. (2004), “Measuring the efficiency of Brazilian post office stores using data envelopment analysis”, International Journal of Operations & Production Management, Vol. 24 No. 10, pp. 1055-78.

[5] Churchill, G. A., Jr., and C. Suprenaut (1982), "An Investigation into the Determinants of Customer Satisfaction," Journal of MarketingResearch, 19 (November), 491-504.

[6] Cronin, J. & Taylor, S. (1992). Measuring service quality: a reexamination and extension, Journal of Marketing, 56(3), 55-68. http://dx.doi.org/10.2307/1252296

[7] Carmen, James M. and Eric Langeard (1980), "Growth Strategies of Service Firms," Strategic Management Journal, 1 (January-March), 7-22.

[8] Fabien, L. (2005). Design and implementation of a service guarantee. Journal of Services Marketing, 19(1), 33-38. http://dx.doi.org/10.1108/08876040510579370

[9] Gouveˆa, M.A., Toledo, G.L. and Filho, L.N.R. (2001), “The prices of mailing services evaluated by companies”, Marketing Intelligence & Planning, Vol. 19 No. 4, pp. 282-94.

[10] Gronroos,Christian(1978), "A Service-Oriented Approach to Marketing of Services," European Journal of Marketing, 12 (no. 8), 588-601. (1982), Strategic Management and Marketing in the Service Sector, Helsingfors: Swedish School of Economics and Business Administration

[11] Lehtinen, Uolevi and Jarmo R. Lehtinen (1982), "Service Quality: A Study of Quality Dimensions," unpublished working paper, Helsinki: Service Management Institute, Finland OY

[12] Lovelock, C.H. & Wirtz, J. (2004). Services Marketing: People, Technology, Strategy (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

[13] McConnell, J. D. (1968), "Effect of Pricing on Perception of Product Quality," Journal of Applied Psychology, 52 (August), 300-303.

[14] Morlok, E.K., Nitzberg, B.F., Balasubramaniam, K. & Sand, M.L. (2000). The parcel service industry in the U.S.: Its size and role in commerce. School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA. Retrieved November 26, 2011 from http://www.seas.upenn.edu/sys/logistics/parcelstudy.html.

[15] Parasuraman, A., Zeithaml, V. A. & Berry, L. L. (1985), A Conceptual Model of Service Quality and Its Implications for Future Research. Journal of Marketing, 49(4), 41 – 50. http://dx.doi.org/10.2307/1251430

[16] Olander, F. (1970), "The Influence of Price on the Consumer's Evaluation of Products," in Pricing Strategy, B. Taylor and G. Wills, eds., Princeton, NJ: Brandon/Systems Press.

[17] Smith, Ruth A. and Michael J. Houston (1982), "Script-Based Evaluations of Satisfaction with Services," in Emerging Perspectives on Services Marketing, L. Berry, G. Shostack, and G. Upah, eds., Chicago: American Marketing, 59-62.

[18] Upah, Gregory D. (1980), "Mass Marketing in Service Retailing: A Review and Synthesis of Major Methods," Jour- nal of Retailing, 56 (Fall), 59-76.

[19] Vijayvargiya. A, & Dey, A.K. (2010). An analytical approach for selection of a logistics provider. Journal of Management History, 48(3), 403 – 418.

[20] Zeithaml, Valarie A. (1981), "How Consumer Evaluation Processes Differ between Goods and Services," in Marketing of Services, J. Donnelly and W. George, eds., Chicago: American Marketing, 186-190.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Appendix B

The International Journal's
Research Journal of Social Science & Management
New ISSN:2251-1571
Old ISSN:2010-457X

A Markov Chain approach to determine the optimal performance period and bad definition for credit scorecard

Murphy Choy,
School of Information System,
SMU, Singapore

Ma Nang Laik,
Assistant Professor,
School of Information System,
SMU, Singapore

ABSTRACT

Performance period determination and bad definition for credit scorecard has been a mix of fortune for the typical data modeler. The lack of literature on these matters led to a proliferation of approaches and techniques to solve the problems. However, the most commonly accepted approach involves subjective interpretations of the performance period and bad definition as well as being chicken and egg problem. These complications result in poorly developed credit scorecard with minimal benefits to the banks. In this paper, we will be recommending a simple and effective approach to resolve these issues.

INTRODUCTION

Credit risk scorecard is an important tool in the tool box of the banking industry. It has been widely used to control consumer credit risk and has been extended to small business credit risk (Anderson, 2005; Thomas et. al. 2002). The earliest credit scorecards were developed by Credit Scoring Consultancies as a way for finance companies to identify risky customers that should not have been given a loan. Due to their proprietary nature (or aptly statistical nature) (Anderson, 2005), few understood the mechanism of the scorecard at the point in time. Early practitioners of credit risk scorecard modeling spent massive amount of time refining the techniques used to build the scorecards. Besides refining the techniques, they spent a lot of time explaining the mechanism and philosophical approach to the finance companies to convince them to use the tool.

As time passes, more and more people understood the mechanism of the credit scorecard and are willing to adopt the model to manage their business. The sudden rise in the consumer credit market directly led to the rise of the credit scorecard industry marking a new milestone in the industry (Lewis, 1992). Many big credit scorecard consultancies were established during this period of expansion such as FICO and Experian which results in the huge disparity in the approaches taken to quantify the risk. This huge disparity results in a major argument about the philosophical aspect of credit scoring and how it should be applied.

At the beginning of the credit scorecard industry, they face strong opposition from a variety of established credit risk practitioner where they adopt conservative credit underwriting process which has been the traditional approach in the field. The main criticism against credit scorecard then was that the variables have very little relation to variables which models them

and that the definition used in the modeling can be rather haphazard and offers little help to finance companies who are trying to manage these risks. This strong opposition is also voiced by some authors (Capon, 1976; Rosenberg et. al., 1994). While there has been much refinement of the credit scoring techniques in the banking world (Eisenbeis, 1977; Eisenbeis, 1978), many criticisms have not been satisfactorily resolved.

With the advent of Basel II, there has been widespread discussion about the definition of a bad account in the context of credit portfolio. The accepted definition for Basel II is any accounts with an ever 90 plus days past due within a performance period of 12 months is considered to be a bad account. This definition is controversial as different financial products behave differently. Some credit products such as mortgage takes a long time to any accounts to satisfy the bad definition while in other cases, the period is too long and most accounts will be considered bad by then (Thomas, 2002; Siddiqi, 2006). Thus proper definition is critical to both proper management of risk as well as operational needs of the banks. In the next section, we will describe the process of defining a bad definition and explore some of existing techniques in evaluating the most optimal combination for defining the performance period as well as the selected bad definition.

DEFINING THE PROBLEM

Credit risk scorecard are designed to measure the probability of an event happening. To be able to measure such events, one must define the event in a manner that is easy to measure and does not confuse with other events that may be a combination of the events. The earliest credit scorecards have extremely simple target events such as predicting whether a customer will becomes ever 30 days past due in the next six months. The improvement in the raw computing power has resulted in ease of building more complicated models which attempts to capture more variations of the bad than what is traditionally used in modeling. Below are some examples of good definitions of 'bad' accounts and contrasting them with complicated and infeasible definitions.

Bad Definition for Modeling	
Good	**Bad**
Ever X+ DPD in 3 Months	2 Times X+ DPD in 4 Months
Ever 30+ DPD in 6 Months	6 Times 30+ DPD in 12 Months
Ever 60+ DPD in 9 Months	2 Times 30+ DPD and 4 Times X+ DPD in 10 Months
Ever 90+ DPD in 12 Months	2 Times Consecutive 30+ DPD in 12 Months

Table 1: Bad definitions

The problem with more complicated bad definition is the difficulty in truly understanding the outcome. Let us contrast the good and bad definitions and use the row 3 definitions from table 1. If you were to ask an analyst what it takes to be a bad customer, the answer will be the definition and you wonder, what about customers who are 1 times 90+ DPD or 3 times 30+ DPD in 10 months? Another possible situation that might arise from this definition is the simplification of the complex definition. The first condition is an extension to the second condition which implies that we can simplify slightly to '2 Times 30+ DPD and 2 Times X+ DPD in 10 Months'. One severe issue with using this type of definition is the time period needed. Given 6 times delinquent in 10 months, the probability of such an event will be very unlikely, resulting a small target population for modeling.

We have examples of good 'bad' definitions but we do not know which definition will meet the requirements of modeling credit default events. Getting a good definition for modeling both in terms of delinquency and performance period will be the focus of the next few sections. We will first discuss about the traditional approach of estimating the performance period and delinquency status for default prediction. Once we have discussed the weakness of the techniques, we will demonstrate the simplicity of the Markov chain approach which solves both problems simultaneously.

EVER DPD CURVES ANALYSIS: PERFORMANCE PERIOD PROBLEM

Determination of the performance period is typically achieved using a type of analysis called ever delinquency curves analysis. This analysis works by analysing the ever delinquency curves trend and attempts to identify the point where the rate of increment in the delinquency rates actually slows. Typically, this is done for several vintages for a particular delinquency. Below is an example of such a chart.

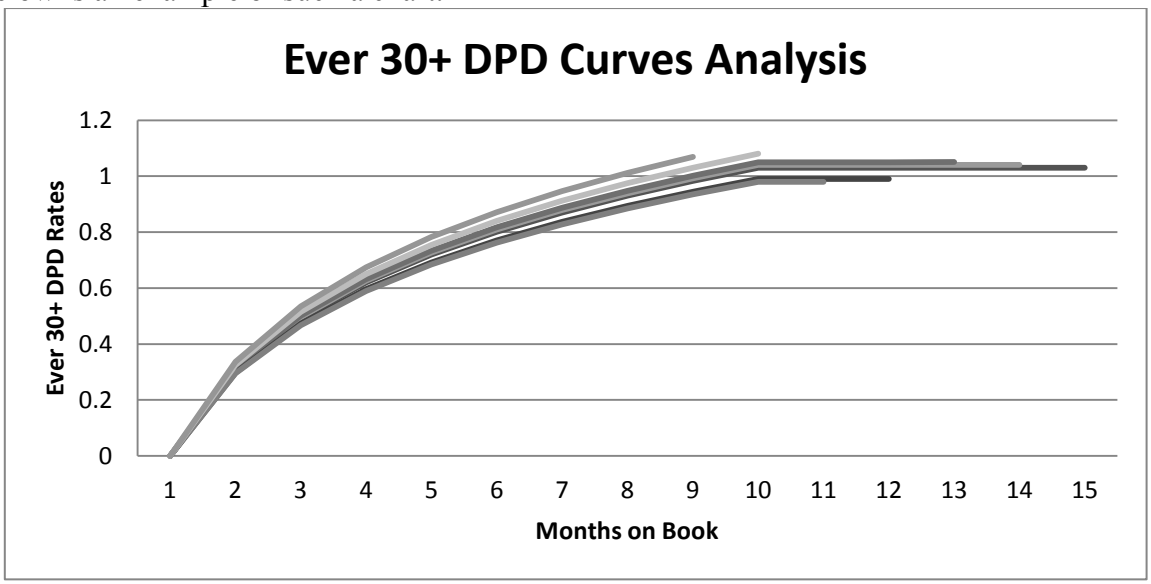


Chart 1: Ever 30+ DPD Trend analysis (Stable leveling)

From the chart 1, we can see the distinct flattening of the ever dpd curves. Being a simulated example, it does not capture the typical unstable flattening of the delinquency curves. Below is another simulated example that looks closer to the ones encountered by analysts in their environtment.

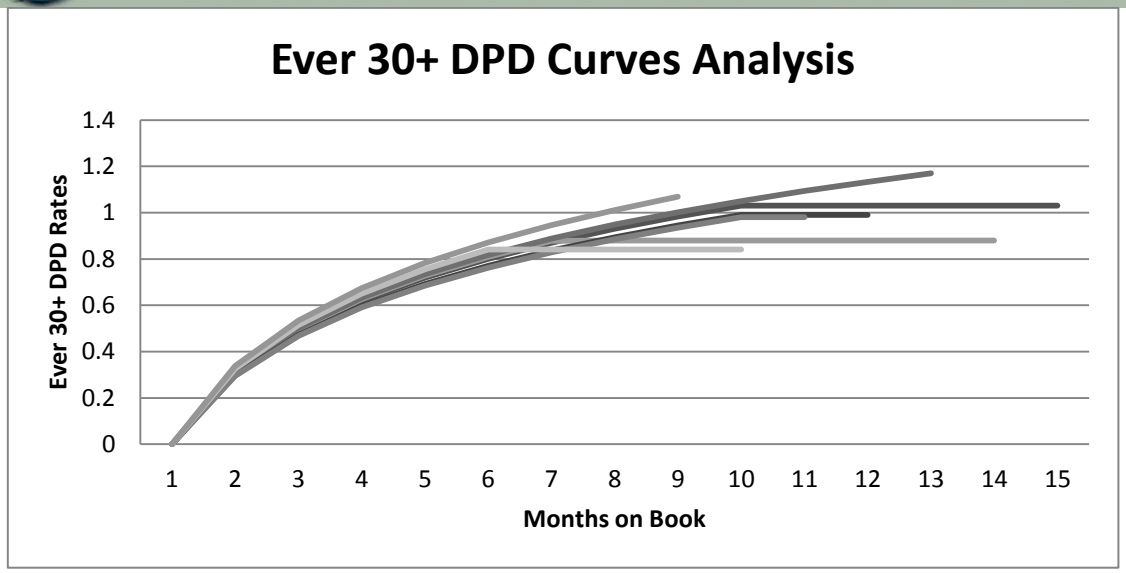


Chart 2: Ever 30+ DPD Trend analysis (Unstable leveling)

From chart 2, we can see that for different vintages, the flattening of the curves are differs from one another and it is extremely difficult to decide on a point in time to identify the start of the flattening. This is compounded by the problem of vintages which are almost ever increasing in their ever bad rate.

The other more serious issue with this analysis is that it requires us to preset the delinquency that will be used for the bad definition to proceed. While multiple iterations will be possible to identify the various optimum performance period for various definitions, it is ultimately a tedious and arduous process.

ROLL RATE ANALYSIS: BAD DEFINITION ISSUES

Once a performance period has been determined, the next important thing to set up is the delinquency definition. Most people would wonder why we do not outright use default or write off as the definition. The reason lies in the fact that outright defaults are small in number and write offs might be manipulated by the management who needs to maintain a good return and low write off portfolio. To compensate for these problems, the most common approach is to define a level of delinquency which signifies the point of no return to default. Usually, any accounts that reach a level of delinquency will have a very high chance of going straight to default. The reason for this is two folds. Firstly, any accounts which have been delinquent for a while will have accumulated massive amount of delinquent amount with interest rate compounding on them. Secondly, if the borrower wishes to repay or possesses the mean to pay, the delinquency will not slip to such a high level of delinquency.

To determine the bad definition, the traditional approach is to use the roll rate analysis (Siddiqi, 2006). Roll rate analysis is a simple Markov Model in which the accounts are grouped according to their ever delinquency status for X months and subsequently whether the account went default in the next Y months. Some variations of the technique exist and one example is the current month vs. next X month delinquency analysis (Siddiqi, 2006). Below is an example of the chart used in the analysis.

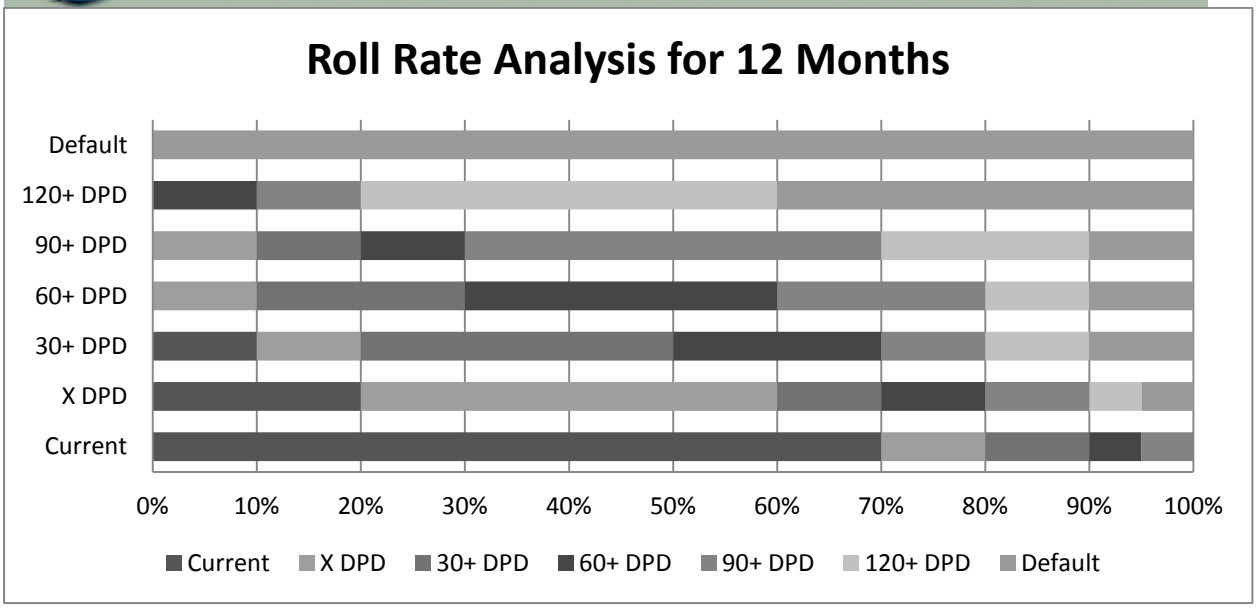


Chart 3:Roll Rate Analysis Chart

From chart 3, we can identify that any accounts going to 120+ DPD in first 12 months have more than 40% chance of going default in the next 12 months. This marks the delinquency status which has a huge group of people going to default once reached. However, as mentioned earlier, the difficulty in executing this analysis is the values used for X and Y in the model. Subjectivity in this case would suggest that there can be multiple definitions used for modeling and that the same chart may result in two different definitions with two analysts.

Together the traditional approaches have their good share of weaknesses which makes them undesirable. In the next section, I will introduce a more robust technique to estimate both delinquency and performance period simultaneously.

MARKOV CHAIN: A PROPERTY THAT SOLVES THE PROBLEM

Markov Chains, also known as transition matrices, are mathematical models which define the probability of an object moving from one state to other states. Depending on the data available, there are several ways to building such a matrix. Below is the mathematical form of the matrix.

States	A1	A2	. . .	A(N-1)	A(N)
A(1)	P(1,1)	P(1,2)	. . .	P(1,N-1)	P(1,N)
A(2)	P(2,1)	P(2,2)	. . .	P(2,N-1)	P(2,N)
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
A(N-1)	P(N-1,1)	P(N-1,2)	. . .	P(N-1,N-1)	P(N-1,N)
A(N)	P(N,1)	P(N,2)	. . .	P(N,N-1)	P(N,N)

Chart 4:Hypothetical Transition Matrix

Each entries in the matrix represent the probability that an object will move to this state given that it starts from the state on the left per turn (usually defined as the time to transit which in this case is one month.). Total sum for each will be 1 for closed systems. One of the interesting property of the Markov chain is that one could calculate the average time spent in each transition states. This calculation is only possible in cases where the matrix contains only transient states (Referring to the case where the row summation does not total to 1). Because of this property, it happens to be uniquely qualified to solve the problem faced in solving the performance period and delinquency to default values.

Let us consider a matrix Q where the states are numbered T = {1,2,…,t} as the set of transient states.

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}} & \cdots & \boldsymbol{P_{1t}} \\ \boldsymbol{P_{i1}} & \boldsymbol{P_{i2}} & \cdots & \boldsymbol{P_{it}} \\ \boldsymbol{P_{t1}} & \boldsymbol{P_{t2}} & \cdots & \boldsymbol{P_{tt}} \end{bmatrix}$$

For each transient state i and j, let m_{ij} denote the expected total number of time periods spent in state j given the starting state of i. Reorganizing the formula yields the following result.

$$\boldsymbol{m_{ij}} = \boldsymbol{\delta(i,j)} + \sum_{\boldsymbol{k}}^{\boldsymbol{t}} \boldsymbol{P_{ik} m_{ik}}$$

$$= \boldsymbol{\delta(i,j)} + \sum_{\boldsymbol{k}}^{\boldsymbol{t}} \boldsymbol{P_{ik} m_{ik}}$$

Where $\delta(i,j) = 1$ when $i = j$ and 0 otherwise. Let M be the matrix containing m_{ij}.

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{m_{11}} & \boldsymbol{m_{12}} & \cdots & \boldsymbol{m_{1t}} \\ \boldsymbol{m_{i1}} & \boldsymbol{m_{i2}} & \cdots & \boldsymbol{m_{it}} \\ \boldsymbol{m_{t1}} & \boldsymbol{m_{t2}} & \cdots & \boldsymbol{m_{tt}} \end{bmatrix}$$

Converting it into the matrix form yields the following equation

$$\boldsymbol{M = I + QM}$$

which can be transformed into

$$\boldsymbol{(I-Q)M = I}$$

and with a little tweak becomes

$$\boldsymbol{M = (I-Q)^{-1}}$$

RESULTS

One important aspect of the data is the required need to filter away customers who have never been delinquent in their entire on account lifetime. This filter is needed as these accounts will artificially increase the mean time spent in current state. At the same time, as mentioned in the earlier sections, we are interested in only accounts that will go to default or write off.

Thus, only accounts that have delinquent history will be useful for us to determine the mean time spent in each state before they reached the point of no return. Let us examine the Markov chain from a credit data set after filtering as shown below.

States	Closed	Current	X	30	60	90	120+ (Write off)
Closed	100%	0%	0%	0%	0%	0%	0%
Current	2%	66%	31%	1%	0%	0%	0%
X	4%	17%	71%	7%	0%	0%	0%
30	4%	4%	15%	45%	30%	3%	0%
60	6%	1%	2%	3%	33%	49%	6%
90	3%	2%	1%	1%	2%	26%	66%
120+ (Write off)	0%	0%	0%	0%	0%	0%	100%

Table 2: Transition Matrix from real data

According to the credit policy, any accounts with 120 days past due are considered as write offs. From the table, we can already observe that any accounts that start in a state of 90 DPD will have more than 50% chance to go to write off. Being the prior state before the final state, it is quite normal to have a higher rate of conversion to the next state. The 60+ DPD state also have very high conversion rate to 90 DPD as well as 120+ DPD. Comparing the conversion rate to the next state to the case of staying or moving to a better state, we can see that people who starts from 60 DPD state has less than 50% chance of staying at 60 DPD or becoming better. Given this case, we can conclude that this is the state which is the point of no return. To attempt to calculate the mean time in state, we will have to first transform the matrix into a canonical form.

States	Current	X	30	60	90	120+	Closed
Current	66%	31%	1%	0%	0%	0%	2%
X	17%	71%	7%	0%	0%	0%	4%
30	4%	15%	45%	30%	3%	0%	4%
60	1%	2%	3%	33%	49%	6%	6%
90	2%	1%	1%	2%	26%	66%	3%
120+	0%	0%	0%	0%	0%	100%	0%
Closed	0%	0%	0%	0%	0%	0%	100%

Table 3:Canonical Form of transition Matrix

Using the transient matrix (shown in table 3 from current to 90 DPD), we can calculate the mean time for each transition state.

States	Current	X	30	60	90
Current	5.9	7.6	1.2	0.6	0.4
X	6.2	10.0	1.7	0.8	0.6
30	1.4	2.5	1.2	0.6	0.5
60	0.9	1.4	0.8	0.9	0.6
90	0.4	0.7	0.2	0.9	0.9

Table 4: Mean Transition Time Matrix

To calculate the mean time taken to reach 60 DPD, all we have to do is to sum the row one up to the point of 60 DPD. From table 4, the result is 5.9 + 7.6 + 1.2 + 0.6 = 15.3 which

approximate to 15 months. Using this information, we can conclude that average time to reach 60 DPD is 15 months and thus the performance period can be set to 15 months

A detailed analysis of the data using the traditional techniques yielded a model with 12 months performance period and 60+ DPD as the bad definition. From this, we can see that the Markov Chain approach produces similar results in a more direct manner.

CONCLUSION

Markov Chain provides Credit Risk analysts with a powerful tool to define their performance period as well as the bad definition that they can use to build credit scorecards on.

REFERENCES

Anderson R.A. (2007). The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk
Management. Oxford University Press: UK.
Capon, N. (1982). Credit scoring systems: a critical analysis. Journal of Marketing 46, 82–91.
Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance and economics. Journal of Finance 32, 875–900.
Eisenbeis, R. A. (1978). Problems in applying discriminant analysis in credit scoring models. Journal of Banking
Lewis EM., An introduction to credit scoring, Athena Press, San Rafael, (1992)
Siddiqi, N. (2006)Credit Risk Scorecards. Wiley.
Thomas, L. C.; Edelman, D. B. and J. Crook (2002) Credit Scoring and Its Applications. SIAM.
Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: a survey. Operations Research 42, 589–613.

***

Appendix C

Uncovering interesting business insights through the use of data analytics in Airport operation: an empirical study

Nang Laik MA,
School of Information System,
Singapore Management University
80 Stamford Road Singapore 178902, Singapore
Email: nlma@smu.edu.sg

Murphy Choy, Michelle Cheong,
School of Information System,
Singapore Management University
80 Stamford Road Singapore 178902, Singapore
Email: murphychoy@smu.edu.sg
Email: michcheong@smu.edu.sg

Abstract
Airport terminals around the world are faced with the limited capacity issue as the number of passengers flowing through the terminals is ever increasing especially for the Asia airports. Many airports in the world have benefited from the increase in their passenger volume by increasing their profitability through the use of shopping malls and duty free shopping. However, any further attempt to increase revenue depends on the capacity of the terminals to accommodate the passengers as well as aircrafts. In this paper, we will focus on the analysis of operational data of an airport and how data analytics can yield interesting insights about the behavior of the airlines as well as the terminals' strategy to manage the airlines. We will also demonstrate how these insights led to a list of proposed solutions which are sufficient to significantly improve the overall performance of the airport and customer satisfaction.

Keywords: Data mining, airport operations, load balancing.

I. Introduction

The global air travel market has been growing at a steady rate for the past decades. Since 2001, the total number of air travel passenger has increased by 25%. This increase translates to approximately 2.25% growth annually in passenger load. This figure has been contributed by the increase in the growth of the industry (refer to Figure 1) that drives the need to improve the infrastructure to meet the needs of the passengers.

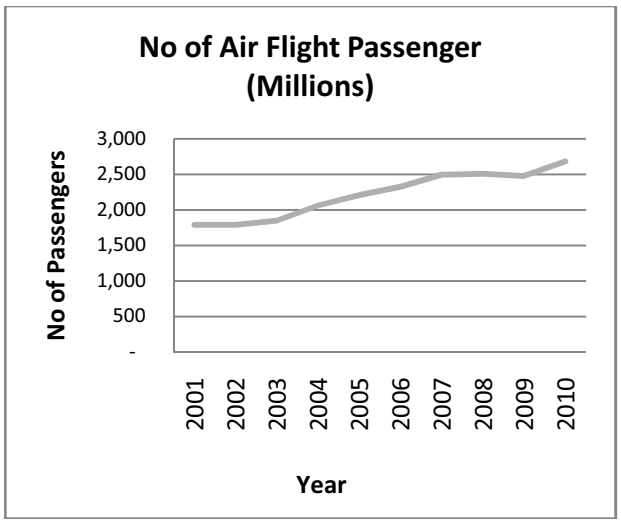


Figure 1: No of Passenger (Worldwide, IATA 2011)

With increasing number of passenger traveling by air, the revenues of the airlines will also increase. From Figure 2 , there has been an increase of 83.33% in revenue from 2001 to 2010. On the other hand, expenses have also increased at almost the same rate of 80% which effectively wipes out the overall profit margin for the airlines. Due to this poor profit margin and high competition between various airport terminals, it is extremely difficult for airport terminals to increase the existing fees that they are

978-0-7695-4770-1/12 $26.00 © 2012 IEEE
DOI 10.1109/SRII.2012.91

charging airlines for the use of terminal and services.

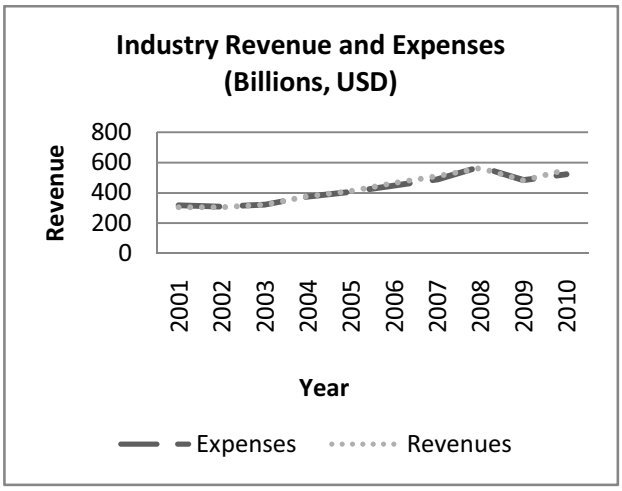


Figure 2: Air Travel Industry Revenue and Expenses (Worldwide, IATA 2011)

There is also another factor which affects the profit margin of the airline. We realized that the breakeven load factor has remained remarkably stable over the years at approximately 65% loaded. This in turn means that, in order to bring a certain number of passengers to an airport terminal, the airlines will have to effectively increase the overall total number of flights to achieve this target of passenger loads.

In the next section, we will discuss the details of the problem in the context of a particular airport in Asia and the actions needed to improve the overall profitability of the business.

II. Problem Description

The main objective for the airport terminal operators is the difficulty in improving the profitability of the terminal business. Most airport operators generate revenue via one of the three following options.

- Airport gate and parking rental
- Check in counter space rental
- Consumer space rental/sales

The first two sources of rental options are highly influenced by the airlines. Thus increasing income from these sources is not possible as most of the airlines are already making losses due to high expenses. Thus, the Asia airport presented in this paper will try to increase income from the third source of rental option. However, there remain challenges in boosting such a rental income.

Passenger satisfaction is a very critical performance indicator for the survival of the airport operations. Being a passenger oriented business; the airport has to maintain excellent standing in customer satisfaction as well as the need to ensure that the passenger will be spending more time and money in the premise to improve the profit.

In this paper, the airport terminal operator is a big player in this industry and operates one of the world's most highly rated airport in Asia. It has turned to data and decision analytics in an attempt to improve the customer satisfaction as well as profitability of the business. They would like to know what insights can analytics provide them and whether any interesting findings which could potentially assist them in making their business more sustainable. This desire to understand their business stems from the drop in passenger satisfaction levels, long waiting time, drop in airport ranking as well as poor profit margin.

In order to increase revenue from consumer space rental/sales, the best way is to increase the human traffic flow through the consumer space. Given that the layout of the airport is fixed, any infrastructure modification will become a major challenge for airport with limited space to maneuver around. The alternative is to increase the overall passenger population which will directly translate into higher human traffic intensity. As mentioned earlier, in order to increase

overall population, we have to increase number of flights which is in turn constrained by the airport infrastructure.

III. Literature Review

Most existing research literatures on airport terminal optimization are classified into two major groups. The first group of literature focused on the issues regarding the air traffic in which the landing and taking off of aircraft are extremely critical [6][9][10][11]. The second group of literature focused on the airport passenger capacity [2][3][4] as well as the configuration of the airport designs [1][7][8] which addressed challenges related to passenger capacity load or flight gate allocation load.

The most common problems mentioned in the journals are the lack of resources that afflicts the terminals' ability to handle flights in and out. Most of the research also focused on the design of the airport to ensure that the arrival and departure of flights are conducted without major hassle. For research works that focused on the passengers, majority of them handled problems related to queuing and luggage handling. However, all the analysis done thus far focused only on the overall operational efficiency without concerns about the revenue generation and the passengers' satisfaction of the airport. At the same time, most of the analyses focused on either on flight load or passenger capacity, and not both. In addition, we note that optimization of either criteria will not be effective and does not offer the optimum solution to the airport problem.

In this paper, we attempt to optimize both flight and passenger loads simultaneously with the objective of improving the passenger satisfaction as well as profitability.

IV. Data Input

From online sources, we extracted the following data of the Asia airport we are working on for analysis.

- Arrival and departure flight information for 2010
- No of Passenger per flight for 2010

At the same time, Table 1 shows the capacities of the three terminals of the airport given as annualized numbers.

Table 1: Terminal Capacity (Annualized)

Terminal	Passenger Capacity	Flight Capacity
1	22,000,000	105,850
2	27,000,000	127,750
3	21,000,000	102,200

A. Analysis of the data

From past experience with time series data, we expected to observe some form of seasonality effect of the passenger throughout the year. However, we noticed that there were no major seasonal effects in the weekly total number of passengers from Figure 3. The three different terminals have curves ranging from 200,000 for the smallest terminal to 280,000 to the most congested terminal with minimal variations. The passenger load for terminal 1 is the highest when we break down the total number of passengers by days of the week in Figure 4. This is a huge disparity given that terminal 2 is the largest terminal compared to the other two terminals (Table 1).

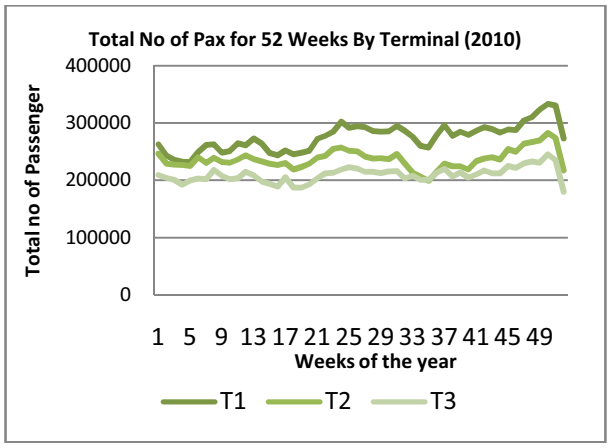


Figure 3: Total No of Passengers for 52 weeks by Terminals

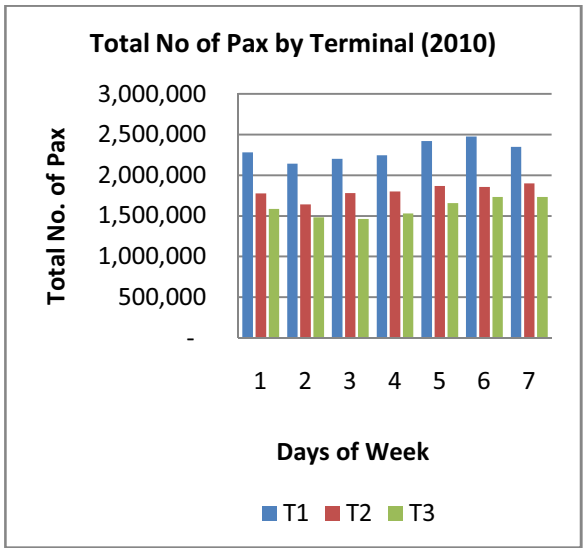


Figure 4: Total Passenger Load by Terminal With respect to Days of week

Close examination of the overall flight load of each terminal for each day of the week for the entire airport pointed us to load imbalance given that the airport terminals were designed to have almost the same capacity. From Figure 5, we can clearly see the impact of overloading of flights in terminal 1 even though the number of passengers handled is relatively close for the different terminals. To better understanding the discrepancy in the report, we have further analyzed the passenger per flight ratio. From Figure 6, we noticed that the passenger per flight ratio for terminal 3 is almost double that of terminal 1.

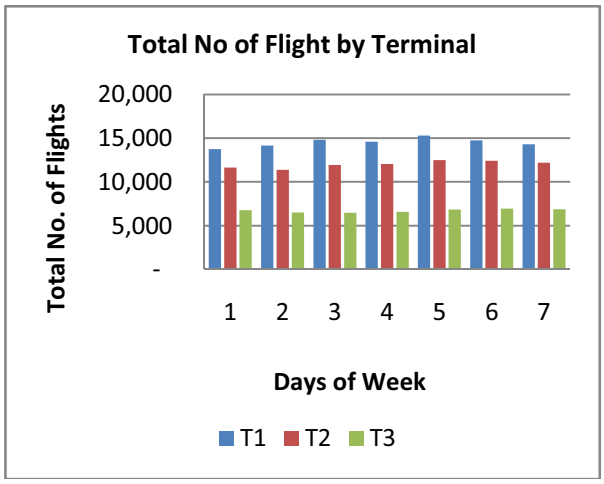


Figure 5: Total Flight Load by Terminal With respect to Days of week

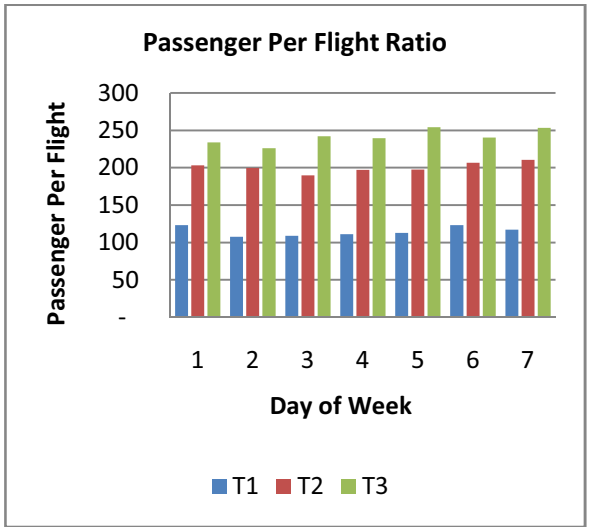


Figure 6: Passenger Per Flight Ratio With respect to Days of week

Given that each terminal has almost similar capacity to handle the flight loads, the huge disparity in handling the flight load was very unusual. The massive difference in the passenger per flight ratio indicated some operational differences between the types of aircraft or route that these terminals handled. Possible reasons raised by the airport management include bigger aircrafts such as the Airbus A380 as well as heavy traffic routes to top major cities. This leads to the next section where issues are identified and the proposed solution is used to rectify the bottleneck and improve the situation.

V. Issues Identified

From the analysis, we have identified several interesting issues with the airline assignment to the terminal.

- Unbalanced flight load utilization for each terminal (90% load for terminal 1 while 60% for terminal 3. See Figure 8.
- High variation in the passenger per flight load. (See Figure 7)

Unbalanced workload at different terminals at the airport means that the resources such as gates or check-in counters are either over-utilized or under-utilized. Flights need the gates in the terminals to load and unload passengers. If the number of flights at the terminal is overloaded, it means that when the flights arrive, the arriving airplane may need to wait a long time on the taxi way to wait for the next available gate. Departing passengers in the congested terminal may also suffer longer waiting time at the check-in counters which leads to lower level of customer satisfaction. Due to the prolonged waiting time at the check-in counters, the passengers may not even have enough time to shop at the retail shops before boarding the aircraft. On the other hand, we have other terminal's resources which are not fully utilized to their maximum capacity and wastage may occur. This is a serious issue for the airport operators to resolve so that the overall profitability and customer satisfaction can be improved.

Most papers in the literature focused on optimizing either the flight load or the passenger load. Their proposed solutions were the optimization of flight load and passenger load independently for each terminal.

To solve the load balancing problem, we have modeled it as a Mixed Integer-Programming Non-Linear (MINLP) model. The main objective of the model is to balance both the flight load and passenger load simultaneously, for all three terminals. This problem is not isolated and only applicable to airport terminal operations. It also exists in various logistic businesses where the load in different warehouses are unevenly distributed causing some sub-par service level performance. A similar problem for container port operations also exists where the management needs to apply some rules to assign the shipping lines to different container terminals. Thus we foresee that our findings and insights provided in this paper will contribute to the applications of data and decision analytics in businesses.

VI. Proposed Solution

Below is our MINLP formulated in a mathematical form.

Subscripts:

i = index for airlines, $1 \leq i \leq I$

j = index for terminals, $1 \leq j \leq J$

Parameter:

A_i - total number of passengers (arrivals, departure and transit) handled by airline i

Z_j - total number of passengers handled by terminal j in a year

K_i - total number of flight (arrival and departure) for airline i in a year

F_j - total number of flights (arrival and departure) assigned to terminal j in a year

P_j - total number of passengers capacity in terminal j

G_j - total number of flight capacity (arrival and departure) in terminal j

B – a group of airlines where they need to be assigned to the same terminal

C – a group of airlines where they need to be assigned to different terminal

Decision variable:

$$x_{ij} = \begin{cases} 1 \text{ if airline i is assigned to terminal j} \\ 0 \text{ otherwise} \end{cases}$$

Object:

$$\min_j (\max\{Var(Z_j), Var(F_j)\})$$

Subject to:

$$\sum_{j=1}^{J} x_{ij} = 1, \forall i \qquad \text{-(1)}$$

$$\sum_{i=1}^{I} (x_{ij} \times A_i) = Z_j, \forall j \qquad \text{-(2)}$$

$$\sum_{i=1}^{I} x_{ij} \times K_i = F_j, \forall j \qquad \text{-(3)}$$

$$Z_j \le P_j, \forall j \qquad \text{-(4)}$$

$$F_j \le G_j, \forall j \qquad \text{-(5)}$$

$$x_{ij} \le M \times x_{i'j}, \forall j, \{(i, i') \in B\} \qquad \text{-(6)}$$

$$x_{ij} + x_{i''j} \le 1, \forall j, \{(i, i'') \in C\} \qquad \text{-(7)}$$

$$x_{ij} \in \{0,1\}$$

The objective function is to minimize the maximum variance between the total passengers and flights load assigned to each terminal.

The first constraint ensures that one airline is assigned to only one terminal to avoid any possible confusion for the passengers. Equations (2-5) ensure that total number of passengers and flights assigned to the terminal do not exceed the terminal capacities. Two important constraints (6-7) also ensure that some airlines have to be assigned to the same terminal due to code sharing among airlines.

A. Preliminary results

The proposed model has been preliminary tested using the publicly available online data which include the flight in and out of the airport for the whole year, current airline-terminal assignment and maximum capacity of the plane based on the aircraft type and specific load factor for the airline. The problem has been solved using the SAS/OR optimization software. The problem was solved within 5 minutes for a realistic problem size of 87 airlines with 232,000 flights and nearly 40 million passengers for the year for 2010. We have plotted the computation output of passenger load and flight load for 1 week (i.e. the flight scheduled are repeated weekly) in Figure 7 and Figure 8.

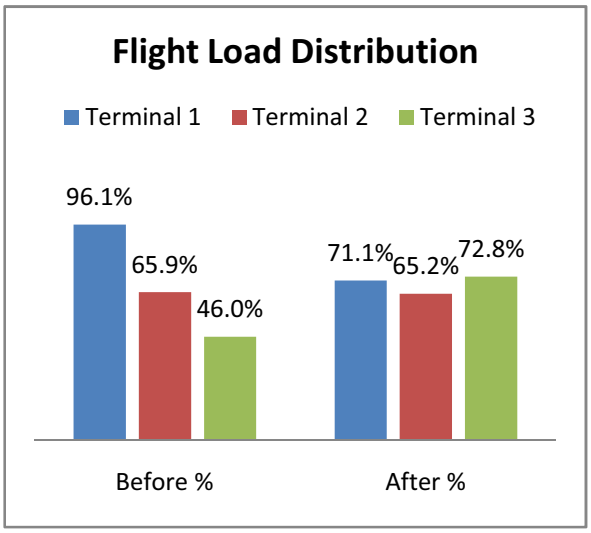


Figure 7: Total Flight Load by Terminal With respect to Days of week Comparison

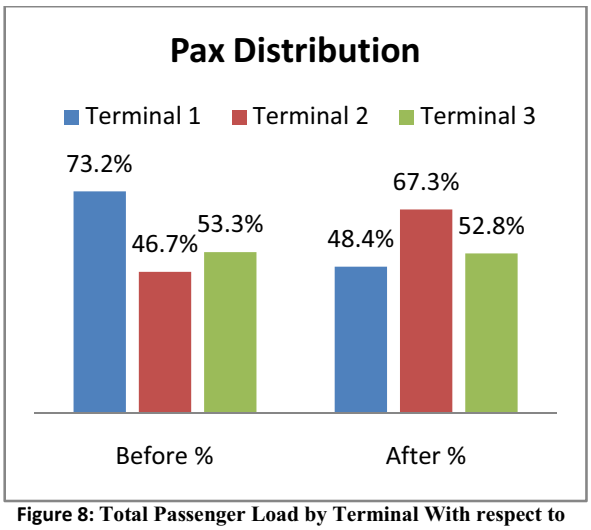


Figure 8: Total Passenger Load by Terminal With respect to Days of week

With the new assignment, we can see that the number of flights handled for each terminal is more evenly distributed with respect to their capacity. The new assignment has also changed the distribution of the passenger to the terminals based on the capacity load. Previously, terminal 1 holds the largest passenger load even though it is not the biggest terminal. The new assignment has made terminal 2 the main handler for the passenger loads shown in Table 2 and Table 3.

Table 2: Terminal Flight Capacity Comparison (Annualized)

Terminal	Before	After	Capacity
1	101,702	75,221	105,850
2	84,187	83,234	127,750
3	46,979	74,413	102,200

Table 3: Terminal Passenger Capacity Comparison Annualized)

Terminal	Before	After	Capacity
1	16,113,471	10,642,188	22,000,000
2	12,619,233	18,177,387	27,000,000
3	11,183,393	11,096,522	21,000,000

A comparison of the flight load has also indicated that the work load is more evenly distributed with respect to their capacity as compared to the past. This indicated that the terminal resources are evenly utilized and the passengers will experience less congestion leading to improves customer satisfaction which will indirectly increase the revenue and profitability of the terminals.

If we use the current airline assignment, terminal 1 is already running at 96% capacity, making it impossible to handle any further increase in the number of passengers. However, with the proposed change, the terminal can handle increase in the number of flights as well as passengers leading to increase in potential revenue while maintaining the flight load and passenger load balance and achieving an acceptable customer satisfaction level.

VII. Conclusion and future research direction

While traditional airport optimization focused on the operational aspect of the terminal management, it does not optimize the overall profitability of the terminal operation which is critical to the viability of the business. In this paper, we have focused on strategic level planning in terminal operation to assign the airlines to three different terminals which belong to the same airport. The objective is to maintain the balanced workload across different terminals at different days of the week and achieving improved customer satisfaction level. Possible future research include combining the micro-level planning of terminal resources (e.g. check-in counters and gates) with the macro –level planning of airport terminals so that the closed feedback loop between the macro-micro level assignments will improve the overall airport terminal efficiency and productivity even further.

REFERENCES

[1]. Braaksma and Cook, 1980. J.P. Braaksma and W.J. Cook , Human orientation in transportation terminals. Transportation Engineering Journal of ASCE 106 TE2 (1980), pp. 189–203.

[2]. Chang et al., 1978a. J.W. Chang, C.D. Le and F. Mangano , Computer simulation of terminal utilization. Airport Forum 8 3 (1978a), pp. 63–67.

[3]. Chang et al., 1978b. J.W. Chang, C.D. Le and F. Mangano , Simulating traffic flows through a terminal. Airport Forum 8 6 (1978b), pp. 79–85.

[4]. Gulewicz and Browne, 1990. V. Gulewicz and J. Browne , Designing an improved international passenger processing facility: A computer simulation analysis approach. In: Airport Terminal and Landside Design and Operation 1990Transportation Research Record 1273, Transportation Research Board, National Research Council, Washington, D.C. (1990), pp. 21–29.
[5]. IATA, 2011
[6]. Le Ny, J., Balakrishnan, H., "Distributed feedback control for an Eulerian model of the National Airspace System", American Control Conference, 2009. ACC '09., On page(s): 2891 - 2897, Volume: Issue: , 10-12 June 2009
[7]. Leihong Li, Clarke, J.-P., "Airport configuration planning with uncertain weather and noise abatement procedures", Digital Avionics Systems Conference (DASC), 2010 IEEE/AIAA 29th, On page(s): 2.B.5-1 - 2.B.5-9, Volume: Issue: , 3-7 Oct. 2010
[8]. McKelvy and Sproule, 1989. F.X. McKelvy and W.J. Sproule , Applications for intraairport transportation systems. In: Airport Landside Planning TechniquesTransportation Research Record 1199, Transportation Research Board, National Research Council, Washington, D.C. (1989), pp. 49–63.
[9]. Ramanujam, V., Balakrishnan, H., "Estimation of arrival-departure capacity tradeoffs in multi-airport systems", Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on, On page(s): 2534 - 2540, Volume: Issue: , 15-18 Dec. 2009
[10]. Sridhar, B., Grabbe, S.R., Mukherjee, A., "Modeling and Optimization in Traffic Flow Management", Proceedings of the IEEE, On page(s): 2060 - 2080, Volume: 96 Issue: 12, Dec. 2008
[11]. Zhao Yifei, "An application of two-tier cooperation problem solution model in Air Traffic Service", Service Systems and Service Management, 2008 International Conference on, On page(s): 1 - 5, Volume: Issue: , June 30 2008-July 2 2008

Appendix D

Metadata of the chapter that will be visualized online

Chapter Title	Data Analysis of Retailer Orders to Improve Order Distribution	
Copyright Year	2015	
Copyright Holder	Springer International Publishing Switzerland	
Corresponding Author	Family Name	**Cheong**
	Particle	
	Given Name	**Michelle L. F.**
	Suffix	
	Division	School of Information Systems
	Organization	Singapore Management University
	Address	80 Stamford Road, Singapore, 178902, Singapore
Corresponding Author	Family Name	**Choy**
	Particle	
	Given Name	**Murphy**
	Suffix	
	Division	School of Information Systems
	Organization	Singapore Management University
	Address	80 Stamford Road, Singapore, 178902, Singapore
	Email	murphychoy@smu.edu.sg
Abstract	Our paper attempts to improve the order distribution for a logistics service provider who accepts order from retailers for fast moving consumer goods. Due to the fluctuations in orders on a day to day basis, the logistics provider will need the maximum number of trucks to cater for the maximum order day, resulting in idle trucks on other days. By performing data analysis of the orders from the retailers, the inventory ordering policy of these retailers can be inferred and new order intervals proposed to smooth out the number of orders, so as to reduce the total number of trucks needed. An average of 20 % reduction of the total number of trips made can be achieved. Complementing the proposed order intervals, the corresponding new proposed order size is computed using moving average from historical order sizes, and shown to satisfy the retailers' capacity constraints within reasonable limits. We have successfully demonstrated how insights can be obtained and new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company.	
Keywords (separated by "-")	Data analytics - Decision analytics - Order distribution - Inventory policy inference	

Chapter 15
Data Analysis of Retailer Orders to Improve Order Distribution

[AU1] [AU2] **Michelle L.F. Cheong and Murphy Choy**

Abstract Our paper attempts to improve the order distribution for a logistics service provider who accepts order from retailers for fast moving consumer goods. Due to the fluctuations in orders on a day to day basis, the logistics provider will need the maximum number of trucks to cater for the maximum order day, resulting in idle trucks on other days. By performing data analysis of the orders from the retailers, the inventory ordering policy of these retailers can be inferred and new order intervals proposed to smooth out the number of orders, so as to reduce the total number of trucks needed. An average of 20 % reduction of the total number of trips made can be achieved. Complementing the proposed order intervals, the corresponding new proposed order size is computed using moving average from historical order sizes, and shown to satisfy the retailers' capacity constraints within reasonable limits. We have successfully demonstrated how insights can be obtained and new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company.

Keywords Data analytics • Decision analytics • Order distribution • Inventory policy inference

15.1 Introduction

Third party logistics companies (3PL) are often faced with the challenges of managing the supply chain efficiency for their clients. For a 3PL who acts as the middle man for the distribution of goods for the brand owner to the retailers, several key performance indices (KPIs) are tracked as part of the service level agreement with their clients. One such KPI is the on-time delivery of orders to the retailers. Late deliveries will affect the sales of the products and may even affect market share of the product.

M.L.F. Cheong (✉) • M. Choy (✉)
School of Information Systems, Singapore Management University,
80 Stamford Road, Singapore 178902, Singapore
e-mail: murphychoy@smu.edu.sg

L.S. Iyer, D.J. Power (eds.), *Reshaping Society through Analytics, Collaboration, and Decision Support*, Annals of Information Systems 18,
DOI 10.1007/978-3-319-11575-7_15

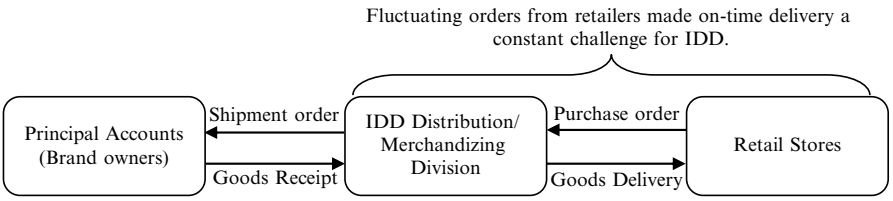


Fig. 15.1 IDD as middle-man for sales & distribution of goods

IDD is a leading integrated distribution and logistics services provider with its headquarter in Hong Kong. IDD provides a full suite of integrated distribution services covering Logistics, Distribution, Manufacturing and International Transportation. The Distribution/Merchandising division plays the middle man role (see Fig. 15.1) in distributing products for their principal accounts (brand owners) to retail stores. Products include food items such as corn flakes and chocolates, and health and beauty items such as toothpaste and shampoo.

The division was often faced with fluctuating orders from the retailers and it did not know how to best manage these fluctuations except to try its best to deliver the orders on time, and face possible penalties from the clients in case of underperforming the contracted KPI. The division wished to understand the fluctuations in orders through analysis of data captured in their IT systems. Through proper data analysis, the division hoped to gain insights on the order behavior of the retailers and propose alternative solution to achieve a win-win situation for the retailers and itself.

15.2 Literature Review

Previous work done on the fulfillment of orders from the upstream supplier or manufacturer to the downstream retailers in a two-stage supply chain under stochastic demand are often focused on sharing of Point-of-Sales (POS) information and implementing Vendor Management Inventory (VMI) so that the supplier can supply the right quantity at the right time to the retailers.

Many papers have highlighted the benefits of information sharing including reduced inventory, daily administration costs and delivery costs. Lee et al. (2000) modeled a two-stage supply chain with one manufacturer and one retailer, to quantify the benefits of information sharing and to identify the drivers that have significant impacts. They showed that manufacturer can obtain larger reductions in average inventory and average cost when the underlying demand is highly correlated over time, highly variable, or when the lead time is long. However, Raghunathan (2001) showed that sharing of demand information is of limited value when the parameters of the demand process are known to both parties, under AR(1) demand with a nonnegative autocorrelation coefficient. The reason is that the manufacturer can forecast the demand information shared by the retailer with a high degree of accuracy using retailer order history, rather than using only the most

recent order from the retailer to forecast the future orders. The accuracy increases monotonically with each subsequent time period. Consequently, the value of information shared by the retailer decreases monotonically with each time period, converging to zero in the limit. Thus, if the manufacturer uses its available information intelligently, there is no need to invest in inter-organizational systems for information sharing purposes.

Yu et al. (2002) also modeled the two-stage supply chain of a beauty product supplier and a retail store. They found that increasing information sharing will lead to Pareto improvement (at least one member in the supply chain is better off and no one is worst off) in the performance of the entire supply chain. Cheng and Wu (2005) extended the two-stage supply chain to consider multiple retailers and allowed correlation of orders to be negative, an extension from Yu et al. (2002). They introduced three different levels of information sharing from level 1 with only knowing retailers' order information; to level 2 with knowing both the retailers' order and customer demand information; and finally to level 3 with real-time information of customer demand through EDI. The optimal inventory policy under each of them was derived. Finally, they showed that both the inventory level and expected cost of the manufacturer decrease with an increase in the level of information sharing. However, they also showed that there was no difference between the inventory level and expected cost of the manufacturer for levels 2 and 3 of information sharing. This implied that there was no need for real-time sharing of demand information or VMI implementation for a two-stage supply chain.

Steckel et al. (2004) stated that whether the sharing of POS information is beneficial or not depends on the nature of the demand pattern represented by the POS information. If the demand pattern conveys continual change in ultimate downstream customer demand, the POS information can in fact distract the upstream decision maker from the more relevant information available from the orders placed by the downstream agent and the supply line. Gaur et al. (2005) extended the results of Raghunathan (2001) to cases in which demand is (AR(p), $p>1$) or (ARMA(p, q), $p>1$, $q>1$). They found that the value of sharing demand information in a supply chain depends on the time series structure of the demand process. When both the demand process and the resulting order process are invertible, demand can be inferred by the manufacturer without requiring further information from the retailer. When demand is invertible but the resulting order process is not, sharing demand information is necessary. They proposed that the demand process is inferable from retailer's order quantity, if the upstream manufacturer's forecast of demand obtained by observing retailer's order quantity, converges almost surely to the actual realization of the demand as time t tends to infinity.

Williams and Waller (2010) compared the order forecasts for the highest echelon in a three-stage supply chain, using POS data versus using order history for cereal, canned soup and yogurt. Their results show that order forecast accuracy depends largely on the product characteristics (seasonal or not) and forecast horizon. In general, POS data produces a better forecast. However, for canned soup which is a seasonal product, POS data did not outperform order history for short term forecasting; whereas and for yogurt which is a short-life span product, POS data performs almost the same as order history.

In our case, IDD did not have any Point-of-Sales (POS) data or shared demand information from the retailers, thus IDD was unable to know or infer the actual demand. Instead, we hope to perform data analysis on historical order information to infer the inventory policies of downstream retailers, and to propose new order intervals and order sizes from historical order data to reduce distribution cost. By playing a proactive role in recommending order interval and the corresponding order size, the retailers need not place order actively, and IDD can better plan distribution to reduce cost.

We could only find two pieces of prior work which have similar objectives like ours to use data analysis to improve supply chain performance. Hausman et al. (1973) analyzed the demand data for 126 women's sportswear over 18 months to obtain three different data-generating processes, (1) ratios of successive forecasts are distributed lognormally; (2) ratios of successive forecasts are distributed as t (Student); and (3) actual demands during unequal time periods are distributed as negative binomial. They concluded that negative binomial was most closely representing the underlying process and simple to adapt to a decision model. Johnston et al. (2003) examined the order size of customers to improve the supply chain. The specific activity mentioned in the paper was that items with intermittent demand, the size of customer orders is required to produce an unbiased estimate of the demand. Also the knowledge of the distribution of demand is important for setting the maximum and minimum stock levels. Both works did not continue to use results of the analysis to make further supply chain related decisions. We think that we are the first to integrate data analytics and decision analytics, where historical data was analyzed to obtain insights to support decision making to improve the supply chain.

Our paper is organized as follow. Section 15.3 will describe the data analysis process to infer the inventory policy of the retailers. Based on the results obtained in Sect. 15.3, we propose a distribution strategy in Sect. 15.4. Based on the proposed distribution strategy in Sects. 15.4, 15.5 and 15.6 will compute the new proposed order interval and order sizes respectively. Section 15.7 aims to assess if the new proposed order sizes will violate retailers' capacity constraint. Section 15.8 compares the number of delivery trips based on the proposed strategy with historical data. Finally, Sect. 15.9 provides the conclusions.

15.3 Data Analysis of Retailer Orders to Infer Inventory Policy

The two sets of data (see Appendix) used for analysis were *Logistic data* and *Store Location data* for a cornflakes product (with each different packaging of the same product represented as a different SKU Code). *Logistic data* provided information on Retailer (identified by CustomerNo), SKU Code, SKU Description, Order Date, Order Quantity, Delivery Date, Delivery Status, Shipped Date, and Shipped Quantity; while *Store Location data* provided the Store Code (identified by Shiptocode), Store Name and Location in geo-information format. In total, there are 326 unique retailers, 191 unique SKU Codes, and 2,681 order records.

With only the historical purchase order information, the initial analysis aimed to categorize the retailers into two possible inventory policies namely, Periodic Review (PR) and Continuous Review (CR). The following assumptions were made:

1. The raw *Logistic data* was reconfigured into a new table with the number of orders for each day of the week (Monday, Tuesday, etc.) for each retailer using Order Date, regardless of the SKU item and order size.
2. Since the objective was to understand the ordering behavior of the retailers, the actual SKU item ordered is immaterial. The analysis result in the appendix supported that the ordering behavior of the retailer was independent of the SKU item ordered.
3. The order size is determined when the retailer has decided to place an order, so it is not the cause for placing order, but rather the result of placing order. Thus, when analyzing the ordering behavior, the order size was not considered. However, the order size would be computed after the order policy and order interval were determined.
4. Without loss of generality, we assumed zero delivery lead time, that is, Delivery Date is the same as Order Date. From the actual data, Delivery Date could be different from Order Date due to planned or unplanned delays.
   (a) Planned delay is usually represented by a fixed delivery lead time T days. As we are only concerned with the delivery of the orders instead of the inventory levels of IDD and the retailers, we can apply the analysis results to positive lead time T by simply shifting the results by T days.
   (b) Unplanned delay is usually due to operational inefficiencies with too many causes, and will not be included as part of the analysis.
5. Only retailers with at least ten orders were included in the analysis to ensure validity of the data analysis.

Based on the assumptions, the data were reconfigured according to day of week j. To explain the data analysis performed, we define the following notations:

- i = Retailer index number, $i = 1$ to I
- j = Day of week corresponding to the calendar date. $j = 1$ to 7, where 1 = Monday, 2 = Tuesday and so on. Note that there might be several orders by the same retailer i on different calendar dates which correspond to the same day of week j.
- $\bar{O}_{ij}$ = Set of orders by retailer i on day of week j
- M_{ij} = Number of orders by retailer i on day of week j, where $M_{ij} = |\bar{O}_{ij}| \geq 0$
- $\bar{R}_i$ = Set of all the orders placed by retailer i.

$$\bar{R}_i = \bar{O}_{i1} \cup \bar{O}_{i2} \cup \bar{O}_{i3} \cdots \cup \bar{O}_{i7} \tag{183}$$

- N_i = Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$
- X_{ij} = Ratio of the number of orders placed by retailer i on day of week j and the total number of orders placed by retailer i.

$$X_{ij} = \frac{\left|\bar{O}_{ij}\right|}{\left|\bar{R}_i\right|} = \frac{M_{ij}}{N_i} \tag{187}$$

- Y_{iw} = Sum of any two ratios X_{ij} of retailer i for any 2 days of week j, where $w = 1$ to ${}^{7}C_{2}$ represents the combination index number and there are ${}^{7}C_{2} = 21$ unique combinations.

The two possible inventory policies considered are:

1. *Periodic Review (PR)* – This policy refers to reviewing the inventory level after a fixed interval period and placing the order quantity sufficient to fill up to the order-up-to level. Usually, small retailers who cannot afford the time and effort to review their inventory on a continuous basis will adopt the Periodic Review Policy. By analyzing the percentage of orders on each day of the week, we could infer the day which the retailer usually placed order.

Rule 1: Periodic Review with Single Dominant Day

If there exist a $Max_j(X_{ij}) > X_{cut}$, *then retailer i is assumed to employ the periodic review policy on the dominant order day j, with a confidence interval of* $(1-\alpha)\%$ *and level of significance of* $\alpha\%$.

In our paper, we have selected X_{cut} = 40 % and state that if there exist a $Max_j(X_{ij}) > 40\%$, then retailer i is assumed to employ the periodic review policy on the dominant order day j, with more than 93.48 % confidence that the observation did not occur by chance with level of significance less than 6.52 %. Refer to Appendix for proof.

Rule 2: Periodic Review on 2 Days, But with Single Dominant Day

If there exist a $Max(Y_{iw}) > Y_{cut}$, *then retailer i is assumed to employ the periodic review policy on 2 days of the week represented by the combination index w, with a confidence interval of* $(1-\alpha)\%$ *and level of significance of* $\alpha\%$. *For this combination w, if* $X_{iq} > X_{ir}$ *where q and r are the days of week represented by combination w, then q will be the dominant order day.*

In our paper, we have selected Y_{cut} = 60 % and state that if there exist a $Max(Y_{iw}) > 60\%$, then retailer i is assumed to employ the periodic review policy on 2 days of the week represented by the combination index w, with more than 97.67 % confidence that the observation did not occur by chance with level of significance less than 2.33 %. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

2. *Continuous Review (CR)* – This policy refers to continuously reviewing the inventory level and order only when the inventory level reaches the reorder point, regardless of the day of week. Usually, larger retailers who have a warehouse and inventory management team can afford to continuously review their inventory and adopt the Continuous Review policy. Similarly, by analyzing the percentage of the total number of orders on each day of the week, we could infer that the retailers who adopted the Continuous Review policy did not have a specific day to place order, so their orders were evenly spread over 7 days.

Figure 15.2 below shows two typical retailers. The blue histogram shows a Periodic Review retailer who placed about 90 % of his orders on Monday, while the red histogram shows a Continuous Review retailer who placed orders evenly on every day of the week.

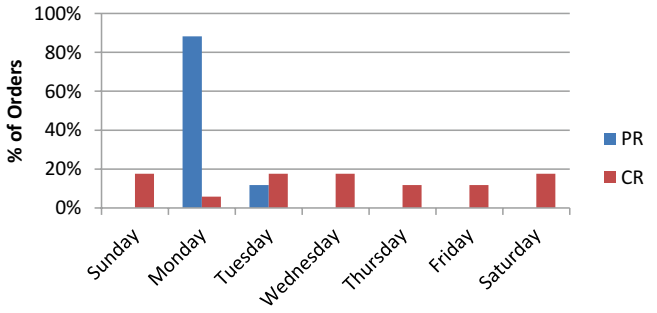


Fig. 15.2 Example of periodic review and continuous review policy retailers

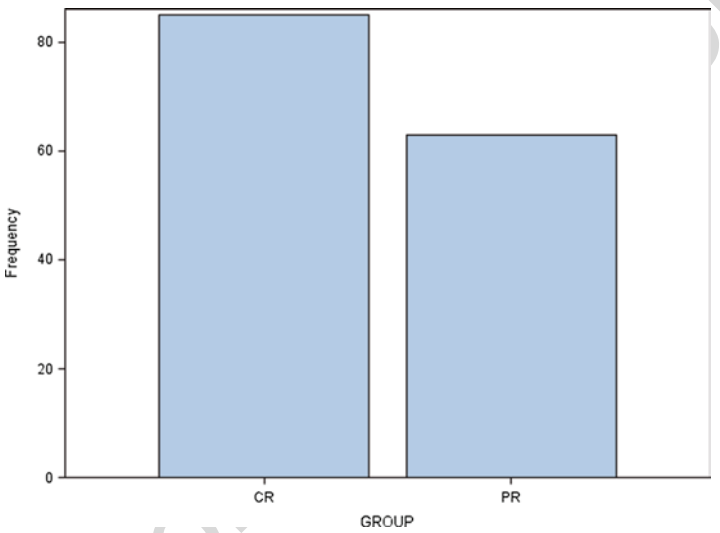


Fig. 15.3 Frequency count of retailers for different inventory policies

To compute the frequency counts for the different number of retailers for each inventory ordering policy, we adopt the following notations:

- $\bar{P}$ = Set of retailers i who employed the periodic review policy based on Rule 1 and Rule 2

$$\bar{P} = \left\{ i \mid \exists Max_j \left(X_{ij} \right) > 0.4 \, or \, Y_{iw} > 0.6 \right\} = \bar{C}'$$

- $\bar{C}$ = Set of retailers i who employed the continuous review policy

$$\bar{C} = \left\{ i \mid \forall Max_j \left(X_{ij} \right) > 0.4 \, or \, Y_{iw} > 0.6 \right\}$$

Our result in Fig. 15.3 shows that most of the retailers employed the Continuous Review policy, that is, $|\bar{C}| > |\bar{P}|$. Since these Continuous Review policy retailers accounted for the bigger portion of the business and orders from them are rather even,

they will form the base load of orders for distribution requiring an almost fixed number of trucks, while the orders from the Periodic Review policy customers will be added on top of the base load, needing the additional trucks.

15.4 Distribution Planning Strategy

After establishing the number of retailer adopting either the Continuous Review (CR) or Periodic Review (PR) policy, we continue to understand how the orders from these retailers distribute across the different days of the week. As every retailer can place order for more than one product, we will define a retailer-product combination since we are only interested to know on which day of the week the retailers place their orders and not what products they order. Each retailer-product combination refers to a particular retailer ordering a particular product. By splitting these retailer-product combination by retailers, Fig. 15.4 shows the distribution for Continuous Review policy retailers (blue bars) which appears to be evenly spread out from Monday to Friday, while the distribution for Periodic Review policy retailers (red bars) has highs and lows from Monday to Friday. This prompted that the fluctuations in orders were caused primarily by the Periodic Review policy retailers. Such fluctuations of orders day to day, will result in needing different number of trucks for each day.

Focusing only on those retailers who adopt the Periodic Review policy, and based on their top order day, Fig. 15.5 shows that the maximum number of orders occurred on Monday, and this number was about twice that of Tuesday, the second highest order day. To ensure on time deliveries on Monday, IDD had no choice but to maintain a large fleet of trucks. However, on the other days of the week (Tuesday, Wednesday, etc.), a smaller number of trucks will be sufficient to complete all deliveries. This will result in excessive number of idle trucks on the other days of the week.

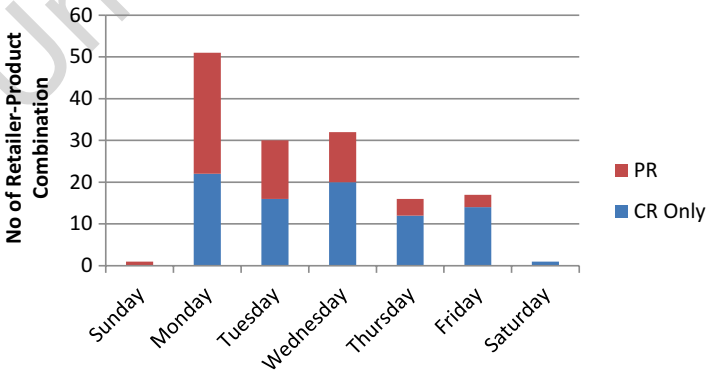


Fig. 15.4 Distribution of retailer-product combination for different day of the week

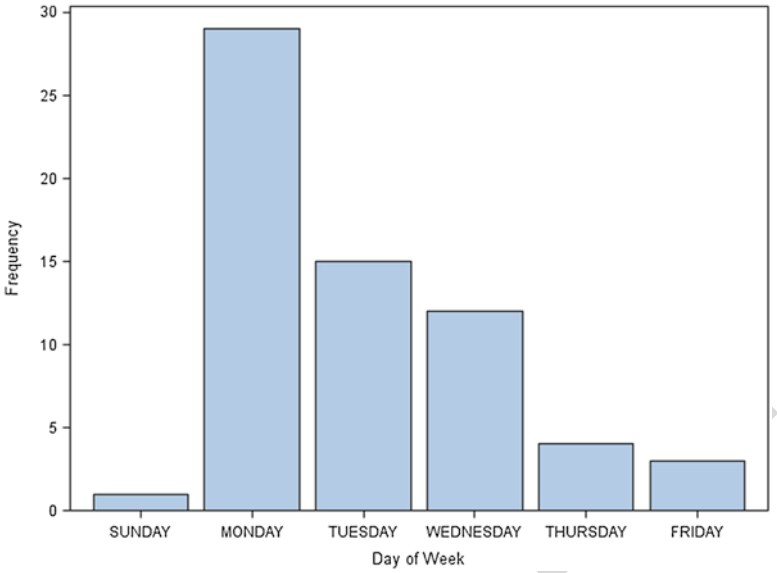


Fig. 15.5 Order frequency for each day of the week for periodic review policy retailers

For the retailers i in set $\bar{P}$,

- $\bar{P}_j$ = Set of retailers i who employed the periodic review policy on dominant day j
- $\bar{P}_1 > \bar{P}_2 > \bar{P}_3 > \bar{P}_4 > \bar{P}_5 > \bar{P}_7$. Note that there are no retailers who employed periodic review policy on Saturday.
- $\bar{P}_1 \sim 2\bar{P}_2$

IDD hoped to even out the distribution for every day of the week, so that the number of trucks used for distribution could be reduced. Since the fluctuations were caused by the Periodic Review policy retailers, the improved distribution plan would only consider smoothing out the orders from these retailers.

IDD can propose to split the retailers for Monday into two groups, each with an order interval of 14 days, instead of 7 days. Group 1 will receive goods on every 1st and 3rd Monday, while Group 2 will receive goods on every 2nd and 4th Monday. The cycle then repeats for 52 weeks in a year. For the other days of the weeks, the retailers will receive goods once a week only on their dominant day.

By carefully allocating retailers belonging to Monday into two groups, IDD can reduce the number of deliveries required for Monday, and thus reducing the total number of trucks required for the entire delivery operations. The allocation of retailers into the two groups (ideally about 50 % of Monday retailers in each group) will depend on their geographical location to minimize the travel distances. Based on the geographical location of the Monday retailers in Fig. 15.6, the Monday PR retailers are divided into five groups in (i) Kowloon, (ii) New World territory region, (iii) Yuen Long & Tuen Mun, (iv) Tung Chung, and (v) the biggest group is in the Hongkong island region. We recommend to split them into two groups,

Fig. 15.6 Geographical location of periodic review policy retailers on Monday

where the biggest group in Hongkong island will be in the first group, while the others will be in the second group, and each group will receive their orders on alternate Monday. Such a split will ensure delivery efficiency.

15.5 Implications of New Proposed Order Interval

Figure 15.7 shows the historical average order interval of Periodic Review policy retailers belonging to Monday. Note that the historical average order intervals are not in multiples of 7 days because these retailers only ordered predominantly on Mondays, but may still order on other days. Our proposed solution was to 'force' them to order only on alternate Mondays, which will make their order interval 14 days. The same principle will apply to retailers who predominantly order on other days of the week, where their average order interval will be 'forced' to be 7 days. This is known as the Power-of-Two principle where by approximating optimal order intervals to the nearest power-of-2 order interval, the total cost is guaranteed to increase not more than 6 %.

Although the total cost to the retailers will not increase by more than 6 %, there are other implications when 'forcing' them to order on alternate Mondays:

- For retailers whose historical average order interval is less than 14 days, they will be receiving orders less frequently than before, and the order size received will be larger. The main concern here would be whether the retailers would have

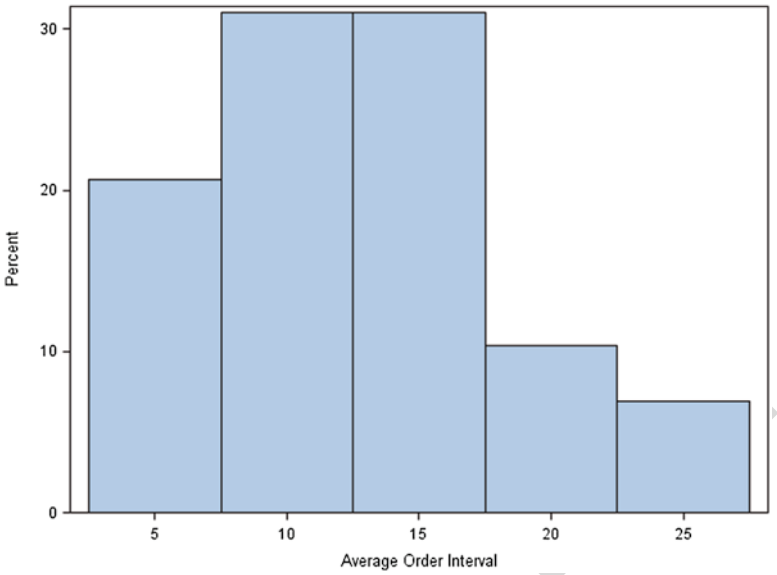


Fig. 15.7 Average order interval for periodic review policy retailers on Monday

sufficient capacity to receive the larger orders. This issue will be addressed in the next two sections.

- For retailers whose historical average order interval is more than 14 days, they will be receiving orders more frequently than before, and the order size received will be smaller. The main concern here would be whether the retailers would have the manpower to receive the orders more frequently. We will not address this issue in this paper.

15.6 Computation for Corresponding Proposed Order Size

The corresponding proposed order sizes can be computed using a moving averaging method, where the averages are computed using historical orders. Assuming historical orders in a particular period will represent future orders in the same period, the proposed order sizes are pre-computed based on historical order data for each retailer, using the proposed order interval of 7 or 14 days.

As defined previously,

- $\bar{O}_{ij}$ = Set of orders by retailer i on day of week j
- M_{ij} = Number of orders by retailer i on day of week j, where $M_{ij} = |\bar{O}_{ij}| \geq 0$
- $\bar{R}_i$ = Set of all the orders placed by retailer i
- N_i = Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$

So, for every retailer i in set $\bar{P}_j$, where j is the dominant order day,

- T_j = Proposed order interval.

$$T_1 = 14,\ T_2 = T_3 = T_4 = T_5 = T_6 = T_7 = 7$$

- k = Order index number of historical orders where $k = 1$ will be the first order. $k = 1$ *to* N_i
- k' = Order index number of proposed orders where $k' = 1$ will be the first order
- O_{ijk} = Order size of historical order k by retailer i for dominant order day j
- t_{ijk} = Time interval between historical order k and order k + 1, by retailer i on dominant order day j. For N_i historical order, there will be $(N_i - 1)$ time intervals.
- $Q_{ijk'}$ = Proposed order size for order k' for retailer i for dominant order day j

The computation method has five main steps for any retailer i with dominant order day j, and proposed order interval T_j.

1. For initialization,

   (a) Compute the first historical average daily demand based on historical order $k = 1$

$$Q_{ij1} = D_{ij1} * T_j$$

   (b) Compute the first proposed order size for order $k' = 1$,

$$Q_{ij1} = D_{ij1} * T_j$$

   This proposed order size Q_{ij1} should cater adequately to demand for the first T_j days.

   (c) Let $D_{ij1} = D_{ijp}$ where the subscript p in D_{ijp} denotes previous average daily demand.

2. Compute a new average daily demand based on the closest equivalent order interval.

$$D_{ijn} = \sum_{k=s}^{s+K} O_{ijk} \Big/ \sum_{k=s}^{s+K} t_{ijk}$$

   Where,

   - s is the starting order index number
   - K is the number of historical orders whose sum of the historical order interval matches closest the proposed time interval T_j K changes for every computation of D_{ijn}.
   - n in D_{ijn} denotes new average daily demand
   - For initialization, $s = 1$. For subsequent iterations, $s = K + 1$.

3. Compute the applied average daily demand by averaging the new average daily demand obtained in step 2, with the previous average daily demand. In case where

the actual demand is known, the actual demand for the past T_j days can replace D_{ijp} for a more accurate average demand to be applied for the next T_j days.

$$D_{ija} = \left(D_{ijp} + D_{ijn}\right)/2$$

4. Compute the adjusted proposed order size for the next T_j days

$$Q_{ija} = D_{ija} * T_j$$

By actively adjusting the proposed order size based on historical value on a moving average, the order size will be able to cater to demand changes.

5. Let $D_{ijp} = D_{ijn}$ and repeat Steps 2, 3 and 4 until the all the proposed order sizes for the entire year of 52 weeks are computed.

Example Computation Based on Table 15.1

1. Initialization

   (a) The first average daily demand was computed from the first order quantity and order interval (i.e. average daily demand = order quantity/order interval). First average daily demand = 10/5 = 2.0

   (b) Using this average daily demand, the proposed order quantity = 14 days* average daily demand = 14* 2.0 = 28. This order quantity should cater adequately to demand for the next 14 days.

2. Compute the new average daily demand based on the closest equivalent order interval. New average daily demand for the closest equivalent order interval of 15 days = (10 + 8 + 9 + 8)/(5 + 4 + 3 + 3) = 2.33
3. Compute the applied average daily demand by averaging the new average daily demand with the previous average daily demand of 2.0. The applied average daily demand = (2.33 + 2.0)/2 = 2.17
4. Adjusted order quantity for the next 14 days interval = 14* 2.17 = 30 (to nearest integer). This new order size of 30 should cater adequately to demand for the next 14 days.
5. Steps 2, 3 and 4 are repeated until the all the proposed order sizes for the entire year of 52 weeks are computed.

Table 15.1 Computation of proposed order size & adjusted order size for 14-day interval

Historical data				14 days order interval	
Order #	Order quantity	Order interval	Average daily demand	Proposed order quantity	Adjusted order quantity
1	10				
2	8	5	2.0	28	
3	9	4	2.0		
4	8	3	3.0		
5	10	3	2.7		30

15.7 Retailers' Capacity Constraint Check

Proposing a longer order interval will result in a larger order size, which may violate the storage capacities at the retail stores. However, the storage capacity at each of the retail stores was not captured in the raw data. We could however infer from the historical purchase order data, assuming that retailers who placed large order in the past would have a large storage capacity.

A measure of reasonableness will be computed as,

$$\text{Ratio Z} = \text{Maximum}\left(\text{Proposed Order Size}\right) / \text{Maximum}\left(\text{Historical Order Sizes}\right)$$

As defined previously,

- $\bar{P}$ = Set of retailers i who employed the periodic review policy
- $\bar{R}_i$ = Set of all the orders placed by retailer i
- N_i = Number of orders by retailer i, where $N_i = |\bar{R}_i| \geq 0$
- k = Order index number of historical orders for retailer i where $k = 1$ to N_i
- O_{ijk} = Order size of historical order k by retailer i for dominant order day j
- k' = Order index number of proposed orders where $k' = 1$ will be the first order
- $Q_{ijk'}$ = Proposed order size for order k' for retailer i for dominant order day j

For every retailer i in set $\bar{P}$, we determine the ratio of $Z_{k'}$ as,

$$Z_{k'} = Max_{k'}\left(Q_{ijk'}\right) / Max_{k}\left(O_{ijk}\right)$$

Table 15.2 shows the percentage of Periodic Review policy retailers with their respective ratio X. Ratio Group 1 has 47 % of the retailers who have Ratio $Z_{k'} < 1$, which means that the proposed order size will not exceed their storage capacity. Ratio Group 2 has 34 % of the retailers who have Ratio $Z_{k'}$ between 1 and 2, which means that the proposed order will be within 1–2 times their maximum order size, which is still reasonable. Ratio Group 3 has the remaining 19 % of the retailers who have Ratio $Z_{k'}$ above 2, which means that the proposed order size have a high chance of exceeding their storage capacity. Cost savings derived from the new distribution strategy can be passed on to these retailers to entice them to accept the new order interval and order size, especially for those in Ratio Group 3.

Table 15.2 Ratio $Z_{k'}$ of proposed order size/ maximum historical order size

Ratio group	%	Ratio $Z_{k'}$
1	47	$Z_k' <= 1$
2	34	$1 < Z_k' <= 2$
3	19	$Z_k' > 2$

15.8 Comparing Number of Delivery Trips

For retailers who employed the Continuous Review policy, there will be no change to the number of orders and thus no change to the number of delivery trips required. For retailers who employed the Periodic Review policy, the number of orders will be changed according to the proposed order intervals (14 days for Monday, and 7 days for other days of the week). The total number of delivery trips made for both policies, was compared with the original number of trips for two groups on Mondays, and 1 group each for Tuesday to Sunday, in Table 15.3.

The number of trips made based on fixed delivery day and fixed interval is reduced by about 20 % and up to 47.3 % for Sunday. The biggest improvement comes from the split of the Monday group into two groups, so that the number of trips needed on any Monday is around 1,100 trips, instead of 2,900 trips in total. This will reduce the total number of trucks required for the entire delivery operations, and in turn reduce the cost of distribution.

15.9 Conclusions

In this paper, we have demonstrated how a logistics company can make use of the data they have captured in their order system to infer the ordering behavior of their retailers. By performing data analysis to categorize the retailers into Periodic Review or Continuous Review policy groups, we could identify that the fluctuations in the number of orders were primarily caused by retailers who employed the Periodic Review policy. These Periodic Review policy retailers were then classified according to their dominant order day and the result showed that the Monday group had double the number of orders than other days of the week. The proposed solution

Table 15.3 Comparison between number of delivery trips

	Monday group 1 (14 day)	Monday group 2 (14 day)	Tuesday (7 day)	Wednesday (7 day)	Thursday (7 day)	Friday (7 day)	Saturday (7 day)	Sunday (7 day)
Original number of trips	1,433	1,455	1,469	1,271	1,074	1,007	67	165
Number of trips based on fixed delivery day and fixed interval	1,148	1,147	1,188	1,034	956	996	67	87
Reduction percentage	19.9 %	21.2 %	19.1 %	18.7 %	11.0 %	1.1 %	0 %	47.3 %

was to split the Monday retailers into two groups with order interval of 14 days, while the other retailers will have order interval of 7 days. The overall reduction in the number of trips made was about 20 % to as high as 47.3 %. The largest savings would be derived from the reduction in the number of trucks to support the entire delivery operations. We have successfully demonstrated how new solutions can be proposed by integrating data analytics with decision analytics, to reduce distribution cost for a logistics company.

15.10 Teaching Note

15.10.1 Overview

[AU3] [AU4]

Many operations management problem ranging from demand forecasting, inventory management, distribution management, capacity planning, workforce scheduling, and queue management are usually solved using known OM/OR techniques such as algorithms, heuristics, and optimization techniques. However, such a typical OM/OR solution methodology often assumes that the actual cause of the problem is known and the problem objective is well defined.

Practitioners like us would know that real business problems do not present themselves clearly, often resulting in people solving the wrong problem. Thus, in this course, the students will be exposed to the Data and Decision Analytics Framework (Fig. 15.8) which helps the analyst to first identify the actual cause of

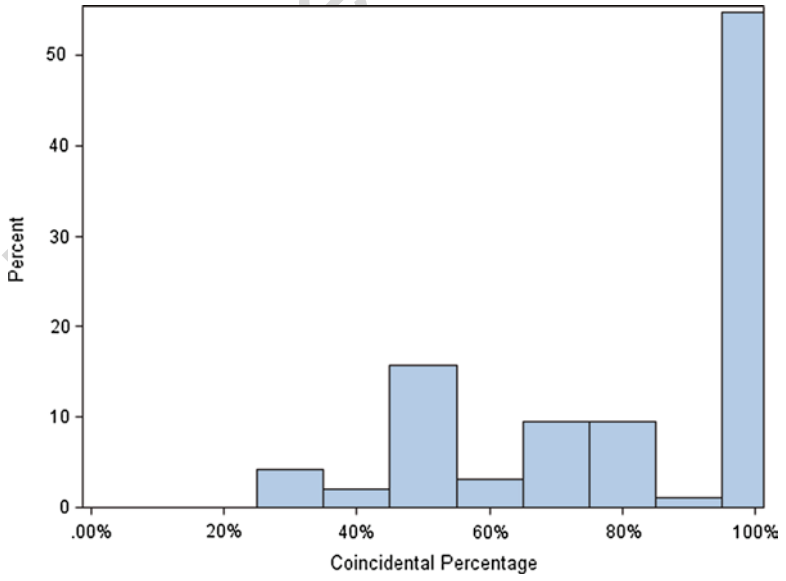


Fig. 15.8 Data & decision analytics framework

business problems by collecting, preparing, and exploring data to gain business insights, before proposing what objectives and solutions can and should be done to solve the problems.

These steps are missing in most problem solving frameworks, particularly in solving operations management problems, where the actual cause of the problem is assumed to be known and the problem objective is assumed to be well defined. However, we advocate that careful data analysis needs to be performed to identify the actual cause of business problems, before embarking on finding the solution.

15.11 Typical Flow of Classroom Activities

A typical flow of classroom activities is depicted in the flow chart in Fig. 15.9. A case usually covers multiple perspectives of operations management topics and the instructor will first cover the topics in terms of the theories and applications. When there are mathematical calculations involved, the instructor can use class activities to supplement and enhance the students' understanding.

After that, the instructor will present the case and facilitate the discussion so that the students can appreciate the case problem and think about the solution methodology according to the Data and Decision Analytics Framework. Once the students understand the intent of the case and what they are supposed to do, the instructor can facilitate the hands-on laboratory session using the step-by-step lab guide. At the end of the lab session, the instructor can instruct the students to complete assignment questions related to the case.

15.12 Introduce Operations Management Topics

For this case, the two topics to be covered include inventory management and distribution management. For inventory management, the understanding of the Periodic Review (PR) policy and Continuous Review (CR) policy should be highlighted. The instructor can ask the students the following questions to facilitate discussions:

- Give examples of goods which the periodic review policy will be more applicable
- Similarly, give examples of goods which the continuous review policy will be more applicable
- What are the advantages and disadvantages of each policy?

For distribution management, the instructor can cover the travelling salesman problem, multiple traveling salesman problem, and vehicle routing problem, introducing the different heuristics which are used to obtain good feasible solution in each problem. The main objective of distribution management is to design tours that will reduce the number of trips made when delivering goods, so as to reduce

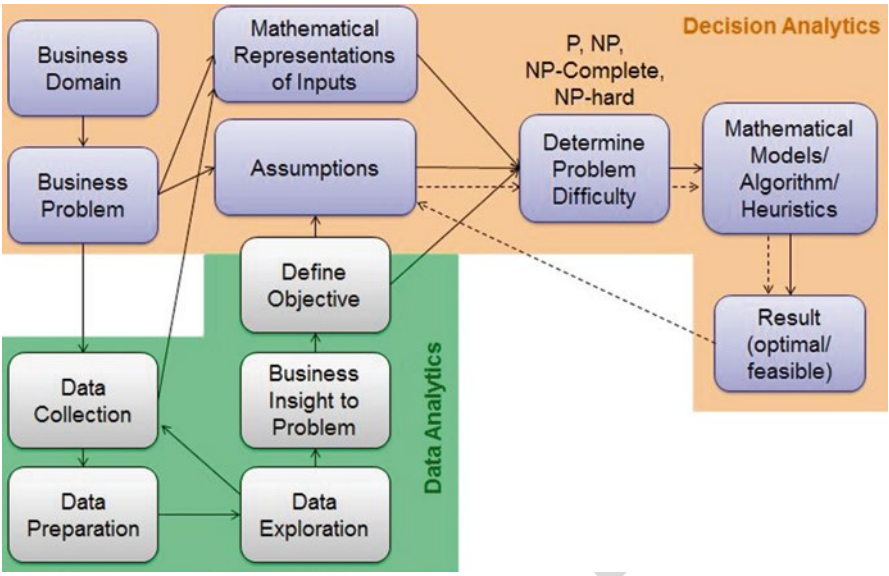


Fig. 15.9 Typical flow of classroom activities

distribution cost. The instructor can ask the students the following questions to facilitate further discussions:

- What other constraints will affect the design of the tour (time window, delivery trucks capacity constraints, client's preferences, traffic conditions)?
- What practical considerations should the vehicle routing planner consider when planning route for a particular driver (familiarity with road, ability to handle different truck size)?
- What practical considerations should the vehicle routing planner consider when planning route for a particular truck (types of goods – refrigerated or not, size of truck, maximum tonnage, door types – open at the back or at the sides)?

15.13 Conduct Case Discussion

15.13.1 Introduce the Case

The case is about IDD which is a leading integrated distribution and logistics services provider with its headquarters in Hong Kong. IDD provides a full suite of integrated distribution services covering Logistics, Distribution, Manufacturing and International Transportation.

The Distribution/Merchandising division of IDD plays the middle man role (see Fig. 15.10) in distributing products for their principal accounts (brand owners) to retail stores. Products include food items such as corn flakes and chocolates,

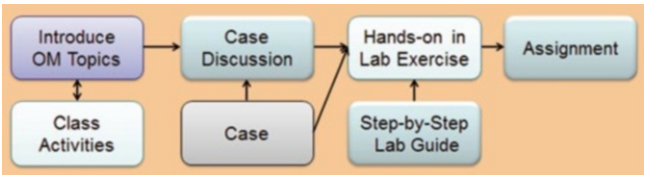


Fig. 15.10 IDD facing problem in distributing fluctuating orders to retailers

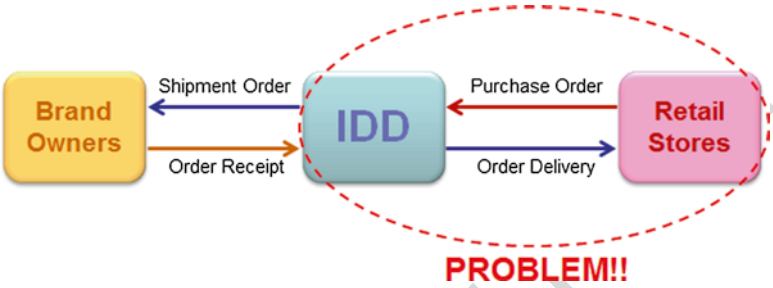


Fig. 15.11 Bullwhip effect experienced in IDD's supply chain

and health and beauty items such as toothpaste and shampoo. The division faces distribution challenges from IDD to the retailers. Orders from retailers fluctuates daily and these fluctuations resulted in the distribution team working very hard with delivery trucks rushing to deliver orders on every Monday, while on other days of the week, the team sees idle trucks parking at the warehouse un-utilized. Playing the passive middle-man role, IDD can only prepare the maximum resource capacities (e.g. drivers and trucks) in order to handle such uncertainties.

The instructor can go further to explain the bullwhip effect in supply chains which is caused by factors such as long lead time, batch ordering and demand variation. In this case, the fluctuations in the retailers orders are likely to be caused by batch ordering behavior of the retailers since demand variation on fast moving consumer goods like cornflakes and toothpaste are relatively low, as shown in Fig. 15.11.

15.13.2 Possible Solutions and Data Provided

After the case introduction, the instruction will ask the students to suggest possible solutions to solve the problem and for each viable suggestion, the students can discuss the pros and cons. One possible suggestion would be to implement Vendor Managed Inventory (VMI) where IDD will deliver the required quantity of products just in time, and the retailers need not place orders actively. For this suggestion, the instructor can ask the students to discuss about the pros and cons of Vendor Managed Inventory.

The Pros include:

- VMI solution will be win-win for both IDD and the retailers
- IDD can plan the deliveries better and reduce the overall cost of deliveries
- The retailers can eliminate manpower to do inventory checks and place orders

The Cons include:

- VMI implementation will require that the retailers share their Point-of-Sales (POS) data with IDD
- Due to confidentiality and trust, most retailers will not be willing to share their POS data

At this point, the instructor can highlight that IDD's IT system stores historical records of the orders from the retailers as well as the store location of each retailer provided in the Appendix of the main paper. With the order data provided (consisting data of 326 retailers and 2,681 orders), the instructor can direct the students to focus on the following four fields:

- Customer No – this is the unique customer ID
- Order Date – this is the order date
- Original Qty – this is the order quantity
- StorerClientCode – this is the store code

With the store location data provided, the instructor can direct the students to focus on the following three fields:

- Latitude – this is the latitude of the store location in geo-information format
- Longitude – this is the longitude of the store location in geo-information format
- Shiptocode – this is the store code which corresponds to StorerClientCode in the Order Data table

15.13.3 Classification Rule

After understanding the data provided, the instructor will lead the discussion on how to infer the retailers' inventory ordering behavior from using the order date. To perform the inference, the instructor needs to explain the Classification Rule (Rule 1 provided in the main paper) which is used to classify the retailers according to Continuous Review (CR) policy or Periodic Review (PR) policy.

At this point, the instructor can ask the students what if X_{cut} is chosen to be say, 60 %? Will the number of retailers categorized into PR retailers be more or fewer?

Upon using the classification rule to categorize the retailers into PR and CR policy, the instructor can explain that by plotting simple bar charts to visualize how

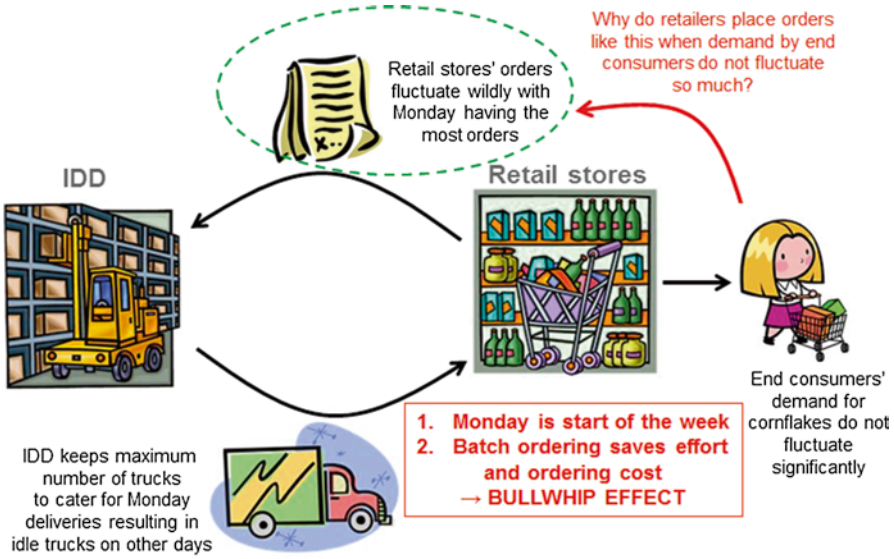


Fig. 15.12 Laboratory exercise activities

many PR and CR retailers place their orders on their dominant order day, students will be able to identify the cause of the order fluctuations and proceed to recommend a new distribution strategy.

15.14 Conduct the Laboratory Exercise

At this point, the students would have appreciated the case problem and understood that they need to perform the laboratory session with the following tasks (depicted in Fig. 15.12),

1. Infer the retailers' inventory ordering behavior by categorizing them into PR and CR according to the classification rule
2. Plot bar charts to visualize the distribution of the retailers according to their dominant order day, and use the bar charts to deduce the root cause of the order fluctuations
3. Propose new distribution strategy which can allow IDD to play a more active role to plan the delivery of the orders to the retailers on each day of the week, and propose the quantity to deliver
4. What constraints must IDD consider and how can IDD ensure that the new distribution strategy is practical (e.g. retailers' capacity challenge)?
5. Justify that the new distribution strategy will result in cost reduction.

15.15 Ensure Learning Outcomes Are Achieved

The entire case aims to achieve several learning outcomes:

1. *Exposure to supply chain business domain covering two major operations management topics including inventory management and distribution management*
   This learning outcome is achieved when the instructor covers the two operations management topics on the theories and the applications, together with class discussion and supplemented with class activities if needed.
2. *Ability to identify the actual cause of business problem by collecting, preparing, and exploring data to gain business insights, before proposing what objectives and solutions can and should be done to solve the problems using the Data and Decision Analytics Framework*
   This learning outcome is achieved when the students apply the steps in the Data & Decision Analytics Framework.
3. *Ability to propose solutions which are practical and provide cost justification*

The third learning outcome is achieved when the students perform the computations for the new proposed order size for the retailers' capacity constraint check and compute the reduction in the number of trips.

Finally, to further enhance the understanding of the case, the students can be asked to complete an assignment with the following question:

Assuming that you can dictate the type of data and information you can get from the business and you can propose a new "Order-to-Distribution-Process", propose an alternative solution to improve distribution and list the types of data needed from new the business process. Map the process flow for your proposed solution.

15.16 Appendix

1. *Logistic Data*

The logistic data contains information about the logistic transport of the goods to the retailer. Here, the retailer is identified by CustomerNo. Table 15.4 also contains some information about the expected delivery of the goods.

2. *Store Location Data*

The store location data contains the information of all the retailers' store location in geo-information format. Here in Table 15.5, the retailer is identified by Shiptocode.

3. *Proofs for Rules 1 & 2*

(a) *Proof for Rule 1: Periodic Review with Single Dominant Day*

If there exist a $Max_j(X_{ij}) > X_{cut}$, *then retailer i is assumed to employ the periodic review policy on the dominant order day j, with a confidence interval of* $(1-\alpha)\%$ *and level of significance of* $\alpha\%$.

Table 15.4 Comparison between number of delivery trips

Field name	Field description
CountryCode	Country Code
CustomerNo	Customer ID
ExpectedDeliveryDate	Expected Delivery Date of Good
OrderDate	Order Date
OrderKey	Order Key
OriginalQty	Original Order Quantity
PODDeliveryDate	Final Delivery Date
PODStatus	Final Delivery Status
PODStatusDescription	Final Delivery Status Description
PrincipalCode	Principal Code
PrincipalDescription	Principal Description
ShippedDate	Shipped Date
ShippedQty	Shipped Quantity
SkuCode	SKU Code
SkuDescription	SKU Description
StorerClientCode	Storer Code

Table 15.5 Comparison between number of delivery trips

Field name	Field description
Latitude	Latitude
Longitude	Longitude
Shiptoaddress1	Address 1
Shiptoaddress2	Address 2
Shiptocity	City
Shiptocode	Store Code
Shiptoname	Store Name
Storerkey	Storer ID
Storername	Storer Name

Consider an order from retailer i which can occur on any of the 7 days of the week.

- The probability of the order falling on a particular day of interest is $\frac{1}{7}$, and we call this the probability of success.
- Thus, the remaining probability of the order *not* falling on that particular day of interest is $\frac{6}{7}$, and we call this the probability of failure.
- This allows us to formulate a Binomial Test with $p = \frac{1}{7}$ and number of trials $= 7$, to determine the X_{cut} with the corresponding confidence interval $(1-\alpha)\%$ and level of significance $\alpha\%$.

From Table 15.6, it is observed that:

– If the percentage of occurrence of orders for a particular day of interest is 14.3 %, we are 73.65 % confident that the observation did not occur by chance with the level of significance of 26.35 %.

t6.1 **Table 15.6** PMF and CDF for binomial test for single day of interest

t6.2 t6.3 t6.4 t6.5 Number of occurrence on a particular day	% of occurrence	Probability Mass Function (PMF) of binomial distribution	Cumulative Distribution Function (CDF) of binomial distribution	1 − CDF = α %
t6.6 0	0 %	0.3399	0.3399	0.6601
t6.7 1	$\frac{1}{7}$=14.3%	0.3966	0.7365	0.2635
t6.8 2	$\frac{2}{7}$= 28.6%	0.1983	0.9348	0.0652
t6.9 3	$\frac{3}{7}$ =42.9%	0.0551	0.9898	0.0102
t6.10 4	$\frac{4}{7}$ =57.1%	0.0092	0.9990	0.0010
t6.11 5	$\frac{5}{7}$ =71.4%	0.0009	0.9999	0.0001
t6.12 6	$\frac{6}{7}$=85.7%	0.0001	1.0000	0.0000
t6.13 7	$\frac{7}{7}$ =100%	Approximately 0	1.0000	0.0000

637 – If the percentage of occurrence of orders for a particular day of interest is
638 28.6 %, we are 93.48 % confident that the observation did not occur by chance
639 with the level of significance of 6.52 %
640 – If the percentage of occurrence of orders for a particular day of interest is
641 42.9 %, we are 98.98% confident that the observation did not occur by chance
642 with the level of significance of 1.02 %
643 – And so on.

644 • In our paper, we have selected X_{cut} =40% and state that if there exist a
645 $Max_j(X_{ij})>40$ %, then retailer i is assumed to employ the periodic review policy
646 on the dominant order day j, with more than 93.48 % confidence that the obser-
647 vation did not occur by chance with level of significance less than 6.52 %.

648 (b) *Proof for Rule 2: Periodic Review on 2 days, but with Single Dominant Day*

649 *If there exist a* $Max(Y_{iw})> Y_{cut}$, *then retailer i is assumed to employ the periodic*
650 *review policy on 2 days of the week represented by the combination index w, with a*
651 *confidence interval of* $(1-\alpha)$% *and level of significance of* α%. *For this combina-*
652 *tion w, if* $X_{iq}>X_{ir}$ *where q and r are the days of week represented by combination w,*
653 *then q will be the dominant order day.*

654 We apply a similar Binomial Test here by grouping the 2 days of interest as 1
655 group, and the remaining 5 days as the other group.

Table 15.7 PMF and CDF for Binomial Test for 2 Days of Interest

Number of occurrence on a particular day	% of occurrence	Probability Mass Function (PMF) of binomial distribution	Cumulative Distribution Function (CDF) of binomial bistribution	1 − CDF = α %
0	0 %	0.0949	0.0949	0.9051
1	$\frac{1}{7}$ =14.3%	0.2656	0.3605	0.6395
2	$\frac{2}{7}$ = 28.6%	0.3187	0.6792	0.3208
3	$\frac{3}{7}$ =42.9%	0.2125	0.8917	0.1083
4	$\frac{4}{7}$ =57.1%	0.0850	0.9767	0.0233
5	$\frac{5}{7}$ =71.4%	0.0204	0.9971	0.0029
6	$\frac{6}{7}$ =85.7%	0.0027	0.9998	0.0002
7	$\frac{7}{7}$ =100%	0.0002	1.0000	0.0000

- The probability of the order falling on two particular days of interest is $\frac{2}{7}$, and we call this the probability of success.
- Thus, the remaining probability of the order not falling on that two particular days of interest is $\frac{5}{7}$, and we call this the probability of failure.
- This allows us to formulate a Binomial Test with $p = \frac{2}{7}$ and number of trials = 7, to determine the Y_{cut} with the corresponding confidence interval $(1-\alpha)$ % and level of significance α %.

From Table 15.7, it is observed that:

  - If the percentage of occurrence of orders for any of the two particular days of interest is 14.3 %, we are 36.05 % confident that the observation did not occur by chance with the level of significance of 63.95 %.
  - If the percentage of occurrence of orders for any of the two particular days of interest is 28.6 %, we are 67.92 % confident that the observation did not occur by chance with the level of significance of 32.08 %
  - If the percentage of occurrence of orders for any of the two particular days of interest is 42.9 %, we are 89.17 % confident that the observation did not occur by chance with the level of significance of 10.83 %
  - And so on.

- In our paper, we have selected $Y_{cut} = 60\%$ and state that if there exist a $Max(Y_{iw}) > 60\%$, then retailer i is assumed to employ the periodic review policy

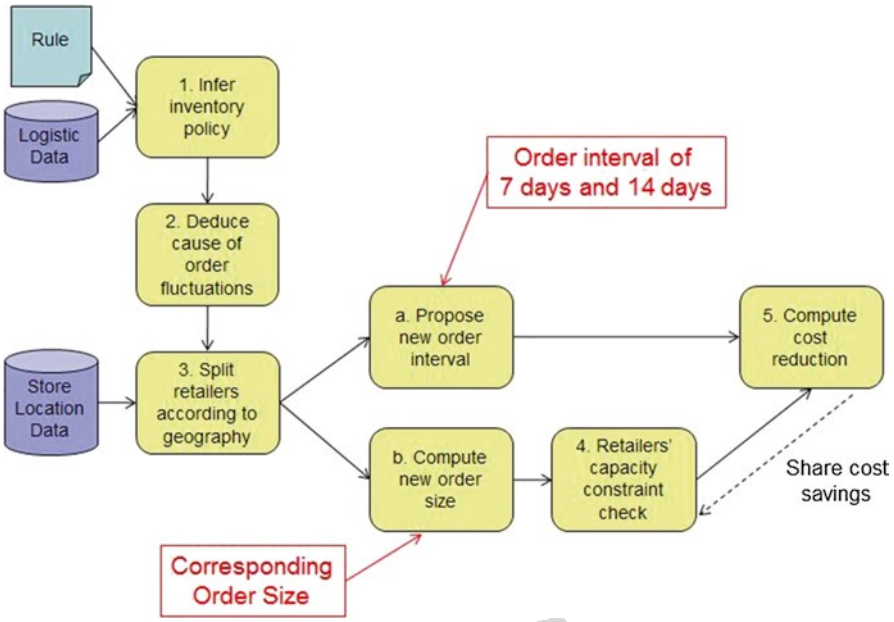


Fig. 15.13 Coincidental analysis of dominant day for periodic review policy retailers

on 2 days of the week represented by the combination index w, with more than 97.67 % confidence that the observation did not occur by chance with level of significance less than 2.33 %. For this combination w, if $X_{iq} > X_{ir}$ where q and r are the days of week represented by combination w, then q will be the dominant order day.

4. *Coincidental Analysis of Ordering Practice for Period*

Further analysis of the Periodic Review policy retailers in the Fig. 15.13 below shows the coincidental analysis of dominant day for retailers. About 60 % of them have 100 % of their orders fixed on the same day of the week. This further justified that the ordering pattern of the Periodic Review policy retailers is independent of the SKU item ordered.

Bibliography

Michelle L. F. Cheong is currently an Associate Professor of Information Systems (Practice) at the School of Information Systems (SIS) at Singapore Management University (SMU). Prior to joining SMU, she had 8 years of industry experience leading teams to develop complex IT systems which were implemented enterprise-wide covering business functions from sales to engineering, inventory management, planning, production, and distribution.

Upon obtaining her Ph.D. degree in Operations Management, she joined SMU in 2005 where she teaches the *Business Modeling with Spreadsheets* course at the undergraduate level and is the co-author of the book of the same name. She also teaches in three different master programmes at SMU on different variants of spreadsheet modeling courses covering different domains, including financial modeling, innovation modeling and IT project management. She recently designed and delivered an *Operations Focused Data Analytics* course for the Master of IT in Business (Analytics) programme at SIS. Apart from her teaching, Michelle is also the Director of Postgraduate Professional Programmes at SIS where she is in charge of two master programmes and Continuing Education & Training.

Michelle has won several awards including the *Most Promising Teacher Award* in 2007 and the *Postgraduate Professional Programme Development Award* in 2013, both from the SMU Center for Teaching Excellence. In addition, she has recently bagged the inaugural *Teradata University Network (TUN) Teaching Innovation Award* 2013, which recognizes excellence in the teaching of Business Intelligence and Business Analytics at the undergraduate, graduate and/or executive education levels.

Murphy Choy is currently a Data Analytics System and Learning Engineer at the School of Information Systems (SIS) at Singapore Management University (SMU). Prior to joining SMU, he had 4 years of industry experience in the area of risk analytics covering Asia Pacific, Middle East, Africa and Latin America. He has spearheaded several analytics initiatives in the bank and has done extensive research work for collections, recoveries and Basel II models. Murphy is especially competent in SAS software and many other analytics software, and also responsible for most of the laboratory exercises designed and taught in the Master of IT in Business (Analytics) programme.

He has served as Chairperson for the Singapore SAS user group and Section Chair for the SAS Global User group. He is pursuing his doctorate degree and his research interest is in the field of Operation Management and Text Mining. He has earned a MSc degree in Finance from University College Dublin and a BSc degree in Statistics from National University of Singapore. Murphy is also a co-author of the paper that won the inaugural *Teradata University Network (TUN) Teaching Innovation Award* 2013.

References

Cheng, T. C. E., & Wu, Y. N. (2005). The impact of information sharing in a two-level supply chain with multiple retailers. *The Journal of the Operational Research Society, 56*(10), 1159–1165.

Gaur, V., Giloni, A., & Seshadri, S. (2005). Information sharing in a supply chain under ARMA demand. *Management Science, 51*(6), 961–969.

Hausman, W. H., & Sides, R. S. G. (1973). Mail-order demands for style goods: Theory and data analysis. *Management Science, 20*(2, Application Series), 191–202.

Johnston, F. R., Boylan, J. E., & Shale, E. A. (2003). An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. *The Journal of the Operational Research Society, 54*(8), 833–837.

737 Lee, H. L., So, K. C., & Tang, C. S. (2000). The value of information sharing in a two-level supply
738 chain. *Management Science, 46*(5), 626–643.
739 Raghunathan, S. (2001). Information sharing in a supply chain: A note on its value when demand
740 is nonstationary. *Management Science, 47*(4), 605–610.
741 Steckel, J. H., Gupta, S., & Banerji, A. (2004). Supply chain decision making: Will shorter cycle
742 times and shared point-of-sale information necessarily help? *Management Science, 50*(4),
743 458–464, Special issue on marketing and operations management interfaces and coordination.
744 Williams, B. D., & Waller, M. A. (2010). Creating order forecasts: Point-of-sale or order history?
745 *Journal of Business Logistics, 31*(2), 231–251.
746 Yu, C. Z., Yan, H., & Cheng, T. C. E. (2002). Modelling the benefits of information sharing-based
747 partnerships in a two-level supply. *The Journal of the Operational Research Society, 53*(4),
748 436–446.

Author Queries

Chapter No.: 15 0002214041

Queries	Details Required	Author's Response
AU1	Please check if author affiliations are okay.	
AU2	Please provide E-mail address for "Michelle LF Cheong"	
AU3	"Teaching Note" section has been merged at end of the Body. Please check if okay.	
AU4	Please check if identified headlevels are okay.	

Effective Listings of Function Stop words for Twitter

Murphy Choy
School of Information System
Singapore Management University
Singapore

Abstract— **Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle to the understanding of the content in the documents. To eliminate the bias effects, most text mining software or approaches make use of stop words list to identify and remove those words. However, the development of such top words list is difficult and inconsistent between textual sources. This problem is further aggravated by sources such as Twitter which are highly repetitive or similar in nature. In this paper, we will be examining the original work using term frequency, inverse document frequency and term adjacency for developing a stop words list for the Twitter data source. We propose a new technique using combinatorial values as an alternative measure to effectively list out stop words.**

Keywords- Stop words; Text mining; RAKE; ELFS; Twitter.

I. INTRODUCTION

Text mining comprises of a series of tasks that includes selection of approach, parameter setting and the creation of a stop word list [14][31]. The creation of a stop word list is often viewed as an essential component of the text mining which requires manual labor and investigations to produce. Stop words lists are rarely investigated and validated compared to the results of the mining process or mining algorithm. The lack of research into stop words list creation resulted in extensive use of pre-existing stop word lists which might not be suitable given the differences in the context of the textual sources. Research in the area has identified the weaknesses of standardized stop words list [3][4][23].

With the spread of social media platforms and adoption of such technologies in business and daily life, social media platforms have become one of the most important forms of communication for internet users and companies. Some companies are using Facebook and Twitter system to provide real time interaction with their customers. These social media platforms are beneficial to companies building consumer brand equity [12]. The platforms also act as low cost effective measures to manage complex relations between companies and consumers. The nature of social media also promotes open and transparent resolution of disputes and allows for greater visibility of the disputes to the senior management. Social Media has also proven to be very effective in communicating news such as the occurrence of earthquakes [25][9] and political office election [21][28].

The enormous amount of textual information from Twitter and social media requires extensive amount of data preparation and analysis to reap any benefits. There are many approaches to analyze the data. However, due to the nature and assumptions of the techniques as well as the huge amount of data collected, the data quality has to be of a very high level of quality in order to be effective [5][13][27]. To improve the quality of textual data, many authors have proposed different techniques to extract an effective stop word list for a particular corpus [22][29]. In the next section, we will focus on the common approaches to the development of stop words list.

II. CURRENT APPROACHES

A stop words list refers a set of terms or words that have no inherent useful information. Stop words create problems in identification of key concepts and words from textual sources when they are not removed due to their overwhelming presence both in terms of frequency as well as occurrence in textual sources. Several authors [30][24][17] have argued for the removal of stop words which make the selection of the useful terms more efficient and reduce the complexity of the term structure. The current literature divides the stop words into explicit stop words and implicit stop words.

The common approach is to manually assemble a stop words list from a list of words. This approach is used by several authors [10] and has proven to be generally applicable to a variety of situation [17]. Even though the generic stop words lists generally achieved high accuracies and robust in nature, customized stop words lists occasionally outperforms especially in technical areas. These customized stop words lists were developed based on the entropy lists or unions of the standard stop lists with entropy lists mixed in [23]. Other authors held the opinion that any words that appear too rarely or were longer than a certain length should be removed [16].

There have been other attempts to use a variety of frequency measures such as term frequency, document frequency or inverse document frequency [15][18]. Each of these measures has proven to be effective in extracting the most common words that appear in the documents. The combination of term frequency with inverse document frequency (TF-IDF) measure was widely quoted by text books and papers [15][29] as the most popular implicit approach for creating a stop words list. There were also attempts in using Entropy approach to calculate the probability of a word being a stop word. [32] In non Anglophone languages, there have successes in using weight Chi Square method in classifying stop words. [33] In Rose et. Al. (2010), the authors proposed a new measure called

the adjacency measure to establish whether a particular word is a stop word or a content word. In the next section, we will examine the algorithm described by Rose et. Al.

III. Rapid Automatic Keyword Extraction Stop word list

In the paper "Automatic keyword extraction from individual documents" by Rose et. Al., the authors describe a process to determine the usefulness of that word in describing the contents. Every word is identified and the word co-occurrences are calculated with a score is calculated for each word. Several scoring techniques based on the degree and frequencies of words were evaluated in the paper. In the paper, Adjacency frequency is defined as the number of times the word occurred adjacent to keywords. Keyword frequency is defined as the number of times the word occurred within keywords. The authors noted that selection by term frequency will increase the likelihood of content-bearing words to be added to the stop words list for a specialized topic that result in removal of critical information words. Rose et. Al. describes the adjacency algorithm as 'intuitive' for words that are adjacent to keywords are less likely to be useful than those that are in it. The authors subsequently tested the algorithm using several standardized documents and found the algorithm to be very effective.

However, there are several issues with the use of the adjacency measure.

1) Adjacency measure first assumes the presence of a keyword in which we can use to determine words that are adjacent. This results in the technique being usable only in the case where keywords are specified. In most textual sources, keywords are not available. In the case of Twitter, while you can use query keywords, it may not be useful for general trend extraction from tweets.

2) Adjacent words might be descriptive words which cannot be found within the keywords. In this case, the measure punishes these words.

3) Adjacency measures assumes multiple keywords in order for the between keywords to be found. This is an unlikely situation given that keywords are likely to single words. This makes it very difficult to be applied to Twitter or documents where the keywords are single words.

Given the restrictive nature of the RAKE stop words list generator, it is very difficult to apply the algorithm to a wide spectrum of text mining problems. In the next section, we will extend on the ideas given in Rose et. al. (2010) and present an effective algorithm in listing functional stop words using the combinatorial counts as measure of information value.

IV. Effective Listings of Functional Stop words using Combinatorial Counts

The authors noted that while the adjacency-within factor cannot be easily computed, the combinatorial factor can be computed easily. The combinatorial factor is defined as the number of unique word combination that can be found in the collection of tweets given a start word. The mathematical form is expressed below.

$$\mathrm{TCF} = \sum_{i=1}^{n} f\left(w_{p,n}, w_{p+1,n}\right) \qquad (1)$$

Where n is the number of tweets, p is the position of the word and w_p is the word in the position p. The function f is the indicator function with the following behavior.

$$f = \begin{cases} 1, \text{where} w_p = w \\ 0, \text{where} w_p \neq w \end{cases} \qquad (2)$$

Where w is the word that is being investigated.

The measure is computationally simple and implementable in a variety of programming languages natively. The combinatorial nature of the measure may not be intuitive. Any words can be linked by a number of words in a language to form meaning combinations. Words designed to convey a precise meaning needs to be linked up in a particular combination for the correct meaning to be conveyed. However, words which are commonly used as bridges in sentences will naturally accumulate a large number of combinations in any collection of documents or tweets. If the collection contains a strong theme or event, the words related will have smaller combinations of words. Theoretically, if there are certain words which are important, the number of combinations should only be one. For example, in any discussion about Linear Algebra, many of the technical terms used will naturally have little variations such as 'Linear Models', 'Complement Set'. This is in contrast to words such as 'in the' and 'that is'.

This measure is an alternative approach to the classical techniques of term-frequency and inverse-document frequency. This approach measures the information value of the word not through the conventional Kullback – Leibler framework but through the combinatorial nature of words. As opposed to measuring the information value of words to establish the stop words, the technique focuses on the extreme number of combinations that most non-meaningful words display to establish stop words. Moreover, the use of combinations allow us to naturally manage both words with high and low occurring frequency which presents a problem for the classical framework of TF*IDF without using transformation.

V. Experimental Setup

To validate the prowess of the measure, we conducted experiments with several techniques commonly used in development of stop word list. For all the experiments conducted, we have selected 9 3-days periods containing tweets with the key word search of 'Earthquake'. Each of this period starts 24 hours before the beginning of an earthquake and last till 48 hours after the occurrence of the earthquake. The reason for selecting 9 different periods and earthquakes is to ensure that the experiments will be as unbiased as possible. The use of query based tweets is to ensure that we have some form of central themes which provides some kind of comparison for the words which are not useful or meaningful. This two conditions enable us to assess the overall performance for the techniques tested effectively and unbiased.

The control factor for this experiment is the Fox's and Manu's stop word list. The choice of having two stop word lists is to double validate the techniques as both stop word lists are commonly used for text mining purposes. At the same time,

both stop word lists have different words which can be useful as a further comparison between the efficacies of the techniques. All the words found in both stop word lists are determined to be stop words in the tweets through human examinations of the tweets using random samples of 1000 unique tweets from each period. For the classical techniques such as term frequency and inverse document frequency, we varied the cutoff thresholds before determining the optimal threshold by calculating the precision of the generated list with the stop list for different range of values. In total, we generated about 10 lists per technique.

Once we have generated the lists, we then compare the list across the different levels of threshold in increasing level of liberty in allowing the word to be considered stop word. Both precision and recall are calculated together with F-measure by comparing the list with the control stop word lists. The technique which consistently outperformed the other techniques will be considered to be the most effective stop word list generator.

VI. Results and Analysis

Using the experimental approach described above, we have generated the various stop words lists and compared their performance at detecting stop words which are listed in the Fox's and Manu's list. In the following sections, we will first compare the various measures and their performance with the Manu's list which is the smaller of the two lists. After the initial comparison, we will then further compare the results using the Fox's list for a second level of validation. The results are plotted with the F-Measures and the threshold levels.

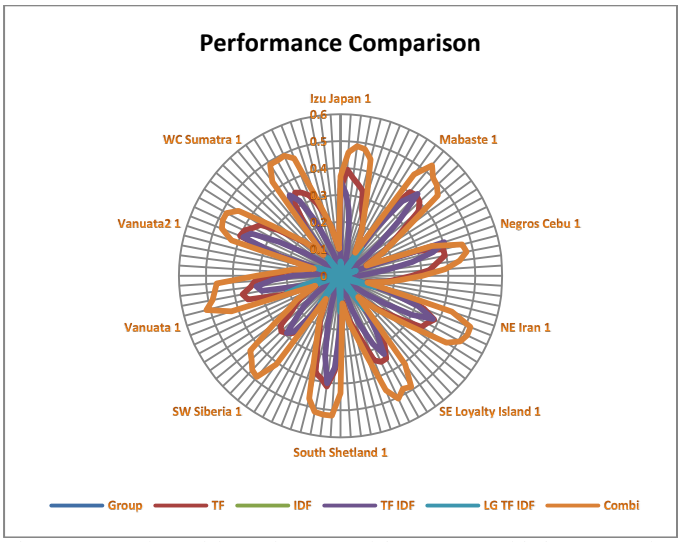


Chart 1: Comparison of the performance of the measures with the Manu's List using Radar Chart

From the chart 1, we can see that the combination technique outperforms most of the other techniques by a fair margin. With the exception of a few initial threshold, where TF*IDF or Log (TF)*IDF variant performs better, the new proposed approach is distinctly better than the other techniques. This superior performance could be attributed to the smaller list of stop words generated by combination approach compared to the other techniques. This effect is further compounded by the small list of stop words in the Manu instance. Many of the words included in the new stop word lists include new words which could be stop words in the context of the Twitter contents.

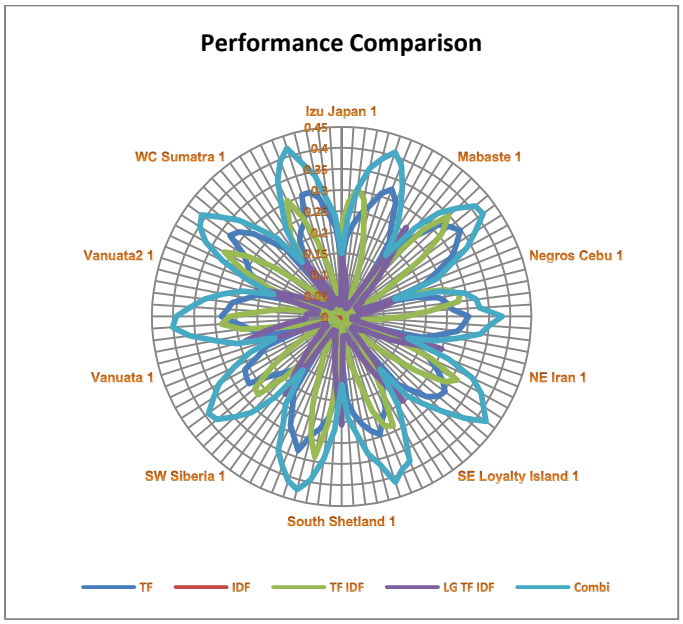


Chart 2: Comparison of the performance of the measures with the Fox's List using Radar Chart

From the chart 2, we can see that the combination technique outperforms most of the other techniques by a fair margin. However, the technique is not as strong as some of the other techniques in the initial threshold levels in some cases as evident in the breaks in the lines of the radar charts. The drop in performance could be attributed to the larger list of stop words covered by Fox's list which is almost three times the size of Manu's list. At the same time, as mentioned earlier, the stop word list generated by the combination technique is also smaller than its TF*IDF and variant counterparts. However, the combination technique still outperforms the other techniques beyond the initial threshold which indicates its superior performance on the overall.

VII. Conclusion

In this paper, we proposed a new method for automatically generating a stop word list for a given collection of tweets. The approach is based on the combinatorial nature of the words in speeches.

We investigated the effectiveness and robustness of the approach by testing it against 9 collections of tweets from different periods. The approach is also compared with the existing approaches using TD*IDF and variants. The results indicated that the new approach is comparable to existing approaches if not better in certain cases.

The direct nature of the combinatorial approach is not normalized and additional research is needed to produce the normalized measure. Other newer approaches such as page-rank approach will also require more research to understand the effectiveness. Future research will also need to investigate the scenario of three or more combinations of words to determine whether they are stop words.

VIII. ACKNOWLEDGEMENT

The author will like to thank the reviewers for their guidance and suggestions. The author will also like to express gratitude to the staffs of School of Information System, Singapore Management University for their guidance and support.

REFERENCES

[1] Busemann, S., Schmeier, S. and Arens, R. G. 2000. Message Classification in the Call Center, Proceedings of the Sixth conference on Applied Natural Language Processing, 158–165.

[2] Blake, C. 2010. Text Mining, ARIST, Vol 45

[3] Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. 1997. Using Taxonomy, Discriminants, and Signatures for Navigating in Text Databases, Proceedings of the 23rd International Conference on Very Large Databases, 446–455.

[4] Chakrabarti, S., Dom, B., Agrawal, R. and Raghavan, P. 1998. Scalabe Feature Selection, Classification and Signature Generation for Organizing Large Text Databases into Hierarchical Topic Taxonomies, The VLDB Journal, Springer-Verlag, 7, 163–178.

[5] Cooley, R., Mobasher, B., and Srivastava, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge Information System, 1-27.

[6] Corbin, J. and Strauss, A. 1990. Grounded Theory Research: Procedures, Canons, and Evaluative Criteria, Qualitative Sociology, 13(1), 3-21.

[7] Corney, M., de Vel, O., Anderson, A., and Mohay, G. 2002. Gender-preferential Text Mining of E-mail Discourse, The 18th annual Computer Security Applications Conference (ACSAC2002).

[8] de Vel, O., Corney, M. and Mohay, G. 2001. Mining E-Mail Content for Author Identification Forensics, SIGMOD Record, ACM Press, 30(4), 55–64.

[9] P. Earle. 2010, Earthquake Twitter. Nature Geoscience, 3:221.

[10] C. Fox. 1992, Lexical analysis and stoplists. In: Information Retrieval - Data Structures & Algorithms, p. 102-130. Prentice- Hall.

[11] Glaser, B., and Strauss, A. 1967. The Discovery of Grounded Theory, Chicago: Aldine Publishing Company.

[12] Jothi,P. et al. 2011. "Analysis of social networking sites: A study on effective communication strategy in developing brand communication", Journal of Media and Communication Studies Vol. 3(7), pp. 234-242, July 2011

[13] Jung, W. 2004. An Investigation of the Impact of Data Quality on Decision Performance, Proceedings of the 2004 International Symposium on Information and Communication Technology (ISICT '04), 166–171.

[14] Keogh, E., Lonardi, S. and Ratanamahatana, C. A. 2004. Towards Parameter-Free Data Mining, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 206-215.

[15] Konchady, Manu. 2007, Text Mining Application programming. Charles River Media Publishing.

[16] Koprinska, I, Poon, J., Clark, J. and Chan, J. 2007. Learning to Classify Email, Information Science, 177, 2167–2187.

[17] Manco, G., Masciari, E., Ruffolo, M. and Tagarelli, A. 2002. Towards An Adaptive Mail Classifier, Proceedings of Italian Association for Artificial Intelligence Workshop.

[18] Mannings, D and Schuetze, H. 1999, Foundation of Statistical Natural Language Processing, MIT Press.

[19] Marwick, A. D. 2001. Knowledge Management Technology, IBM Systems Journal.

[20] Moreale, E. and Watt, S. 2002. Organisational Information Management and Knowledge Discovery in Email within Mailing Lists, In H. Yin et al. (Eds.), Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, 2412/2002, 217-224, Berlin / Heidelberg:Springer-Verlag.

[21] Mungui-Pippidi, Alina and Munteanu, Igor. "Moldova's 'Twitter Revolution.'" Journal of Democracy 20/3 (July 2009): 136-142.

[22] Salton, G. 1971. The SMART Retrieval System—Experiments in Automatic Document Processing, Upper Saddle River, NJ, USA: Prentice-Hall, Inc..

[23] Silva, C and Ribeiro, B. 2003. The Importance of Stop Word Removal on Recall Values in Text Categorization, Proceedings of the International Joint Conference on Neural Networks, 3, 1661-1666.

[24] Sinka, M. P., and Come D. W. 2003. Evolving Better Stoplists for Document Clustering and Web Intelligence, Proceedings of the 3rd Hybrid Intelligent Systems Conference, Australia, IOS Press.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In WWW '10: Proc. of the 19th international Conf. on World wide web, pages 851–860, New York, NY, USA, 2010. ACM.

[26] Tang, J., Li, H., Cao, Y. and Tang, Z. 2005. Email Data Cleaning, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, 2005, 489–498.

[27] Tayi, G. K. and Ballou, D. P. 1998. Examining Data Quality, Communications of the ACM, ACM Press, 41(2), 54–57.

[28] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proc. 4Th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM).

[29] Rose, S., D. Engel, N. Cramer, and W. Cowley. 2010. Automatic keyword extraction from individual documents. In M. W. Berry and J. Kogan (Eds.), Text Mining: Applications and Theory. John Wiley and Sons, Ltd.

[30] Van Rijsbergen, C. J. 1979. Information Retrieval, Newton, MA: Butterworth- Heinemann.

[31] Xu, R. and Wunsch, D. 2005. Survey of Clustering Algorithms, IEEE Transactions on Neural Networks, 16(3), 645-678.

[32] Z. Yao, and C. Ze-wen, "Research on the construction and filter method of stop-word list in text Preprocessing", Fourth International Conference on Intelligent Computation Technology and Automation, 2011.

[33] F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, "Automatic Construction of Chinese Stop Word List", Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015).

AUTHORS PROFILE

Murphy Choy is an instructor with School of Information System at Singapore Management University. He received his MSc finance from University College Dublin, Ireland. He has published papers in risk management, text analytics and operation research. His research interest is in the application of data mining and operation research to real life problem.

Detection of manipulation of inter-bank overnight rate using Euclidean-based time series cluster analysis

Murphy Choy* and Enoch Ch'ng

School of Information Systems,
Singapore Management University,
80 Stamford Road, 178902 Singapore
E-mail: murphychoy@smu.edu.sg
E-mail: enochchng@smu.edu.sg
*Corresponding author

Abstract: The interbank offered rate (IBOR) is the interest rate at which banks can borrow funds from other banks in the interbank market. It is also used as the benchmark upon which rates or financial contracts for less preferred borrowers are based. In light of the recent London IBOR (LIBOR) manipulation incident, this paper seeks to address the concern that IBOR is entirely controlled by the banks. The paper focuses on the comparison between LIBOR and Singapore IBOR (SIBOR) especially with regards to the behaviour of the interest rate with time. The nature of IBORs is such that the rates submitted by the banks will naturally be similar and should not differ excessively from the market as well as the other banks. We will compare the LIBOR and SIBOR from 2005 to 2011 with respect to the one-month rates on an annual basis. The results of our study support that the SIBOR is not manipulated like LIBOR.

Keywords: London interbank offered rate; LIBOR; Singapore interbank offered rate; SIBOR; time series; cluster analysis; Euclidean distance.

Reference to this paper should be made as follows: Choy, M. and Ch'ng, E. (2014) 'Detection of manipulation of inter-bank overnight rate using Euclidean-based time series cluster analysis', *Int. J. Process Management and Benchmarking*, Vol. 4, No. 2, pp.198–212.

Biographical notes: Murphy Choy is currently an Instructor at the School of Information Systems (SIS) at Singapore Management University (SMU). He has served as the Chairperson for the Singapore SAS user group and Section Chair for the SAS Global User group. He has earned his MSc in Finance from the University College Dublin and his BSc in Statistics from the National University of Singapore. He is pursuing his doctorate degree and his research interest is in the field of operation management and text mining.

Enoch Ch'ng is an Associate Professor of Information Systems (Education) at Singapore Management University. He is also the Director of the School of Information Systems (SIS) programmes in Financial Services. Notably, he was a Managing Director in DBS Bank Corporate Office. He was the Executive Director of the Specialist Risk Supervision Department in Monetary Authority of Singapore. He also spent ten years in various roles in UBS. He received his Master's in Aeronautics and Astronautics from the Massachusetts Institute of Technology. He also received his Bachelor's in Mechanical and Aerospace Engineering from Princeton University.

1 Introduction

The London interbank offered rate (LIBOR) scandal first came to light in 2008 at the height of the 2007–2009 global financial crisis. The *Wall Street Journal* first released a study indicating that the banks were deliberately understating the rates in an attempt to improve their financial positions. Both the Bank for International Settlements (BIS) and the British Bankers Association (BBA) responded with statements indicating the reliability of the rates quoted. They cited difficulties in the financial markets as a reason for the discrepancies in the published rates. This position is further supported by the International Monetary Fund's (IMF's) regular reports. However, an independent group of researchers Snider and Youle (2010) did find results which corroborated with the *Wall Street Journal* article. However, the researchers believe that the banks were attempting to profit from the movements in the rates rather than strengthening the banks' positions.

With the new results, financial and fraud investigation commenced on several international banks including Barclays which is the first bank to admit to LIBOR fixing. This investigation rapidly spreads to other financial markets such as Canada and Singapore. Many issues and concerns were raised by the financial regulators about the importance of separation of powers and whether new cross-boundary regulations should be implemented to prevent a repeat of the incident. Others seek a complete overhaul of the interbank offered rate (IBOR) calculation framework in hope that it will prevent manipulations. In the following section, we will demonstrate how the cluster analysis of time series can be used to detect such irregularities and how the IBORs are currently affected by the manipulations.

In the next section, we will review the literature on time series clustering techniques, approaches as well as the pre-processing requirements on the data. There will be deep discussions on the distance calculation which is critical to the clustering algorithm. We survey the various clustering algorithms and present our findings with case studies in different fields of application. We will also examine the relevant approaches and key techniques used in the subsequent sections.

2 Literature review

A key component in clustering is the measure determining the similarity between two set data being compared. The data need neither be compared within the same time frame nor time scale. These data could be captured in various structure and format which include raw values of equal or unequal length, vectors of feature-value pairs and transition matrices among other formats. We will first consider the simplest distance measures, the L_2 measures.

Let x_i and v_j each be a P-dimensional vector. The Euclidean distance is computed as:

$$d_E = \sqrt{\sum_{k=1}^{P} \left(x_{ik} - v_{jk}\right)^2} \tag{1}$$

The root mean square distance (or average geometric distance) is simply:

$$d_{rms} = \frac{d_E}{n} \tag{2}$$

Minkowski distance is a generalisation of Euclidean distance, which is defined as:

$$d_M = \sqrt[q]{\sum_{k=1}^{P} \left(x_{ik} - v_{jk}\right)^q} \tag{3}$$

In equation (3), q is a positive integer. A normalised version can be defined if the measured values are normalised via division by the maximum value in the sequence. For previous examples, the measurement space is primarily distance measures. Clustering algorithms work on both distance and correlation measures. Let x_i and v_j each be a P-dimensional vector. Pearson's correlation factor between x_i and v_j, $Corr_{Pearson}$ is defined as:

$$Corr_{Pearson} = \sum_{k=1}^{P} \frac{\left(x_{ik} - \mu_{x_{ik}}\right)\left(v_{jk} - \mu_{v_{jk}}\right)}{S_{x_i} S_{v_j}} \tag{4}$$

where μ_{xi} and S_{xi} are respectively the mean and scatter of xi, computed as below:

$$\mu_{xi} = \frac{1}{p} \sum_{k=1}^{P} x_{ik} \tag{5}$$

$$S_{xi} = \left[\sum_{k=1}^{P} x_{ik} - \mu_{xi}\right]^{0.5} \tag{6}$$

Assuming that each time series is a piecewise linear function, Möller-Levet et al. (2003) discussed the concept of short time series (STS) distance as the sum of the squared differences of the slopes in two time series being compared. The key objective is to define a distance capable of capturing differences in the shapes, shaped by the relative change of expression and the corresponding temporal information, disregarding the difference in absolute values. Mathematically, the STS distance between two time series x_i and v_j is defined as:

$$d_{STS} = \sqrt{\sum_{k=1}^{P} \left(\frac{v_{j(k+1)} - v_{jk}}{t_{k+1} - t_k} - \frac{x_{i(k+1)} - x_{ik}}{t_{k+1} - t_k} \right)^2} \tag{7}$$

where t_k is the time point for data point x_{ik} and v_{jk}. To remove the effects of different scales, z standardisation of the series is recommended.

Dynamic time warping (DTW) is a form of generalisation of classical algorithms for the purpose of comparing discrete sequences to sequences of continuous values. The approach allows for comparison of separate times series with different length and period. The key objective is to seek the common ground between time series. Given two time series, $Q = q_1,q_2,\ldots,q_i,\ldots,q_n$ and $R = r_1,r_2,\ldots,r_j,\ldots,r_m$, DTW will attempt to align the two series so that their difference is minimised. To this end, an $n \times m$ matrix where the (i, j) element of the matrix contains the distance $d\ (q_i, r_j)$ between two points q_i and r_j. The Euclidean distance is used in most circumstances. A warping path, $W = w_1, w_2,\ldots,w_k,\ldots,w_K$ where $max(m, n) \leqslant K \leqslant m + n - 1$, is a set of matrix elements that must satisfies three constraints which are boundary condition, continuity and monotonicity. The boundary condition constraint requires the warping path to start and finish in diagonally opposite corner cells of the matrix. That is $w_1 = (1, 1)$ and $W_K = (m, n)$. The continuity constraint restricts the allowable steps to adjacent cells. The monotonicity

constraint forces the points in the warping path to be monotonically spaced in time. The warping path that has the minimum distance between the two series is of interest. Mathematically:

$$d_{DTW} = \min\left(\frac{\sum_{k=1}^{K} w_k}{K}\right) \tag{8}$$

Dynamic programming can be used to effectively determine the path. This is achieved by evaluating the following recurrence, which defines the cumulative distance as the sum of the distance of the current element and the minimum of the cumulative distances of the adjacent elements:

$$d_{cum}(i,j) = d(q_i, r_j) + \min(d_{cum}(i-1, j-1), d_{cum}(i-1, j), d_{cum}(i, j-1)) \tag{9}$$

While time series data have time as an additional dimension, performing clustering on time series is similar to clustering on static data. The technique employed to form clusters will vary subject to both types of data available as well as the purpose and application of the analysis. Given the nature of time series, most time series data are continuous values and univariate in nature.

While there are specific algorithms developed solely for the purpose of clustering time series data of different nature, they are similar in their approach of modifying existing non-time series data algorithms to handle time series data. This approach involves raw time series data otherwise also known as raw-data-based approach. It normally involves data pre-processing to convert the time series data to a normal static form. Once the data has been pre-processed, the classical static data clustering algorithms can then be applied.

For clustering multivariate time varying data, Košmelj and Batagelj (1990) modified the relocation clustering procedure that was originally developed for static data. To measure the dissimilarity between trajectories required, they used a cross-sectional approach-based general model that incorporated the time dimension which further developed a specific model based on the compound interest idea to determine the time-dependent linear weights. The cross-sectional procedure ignores the correlations between the variables over time and works only with time series of equal length. The best clustering among all the possible clustering is the one evaluated with the minimum generalised Ward criterion function. Using the same cross-sectional approach, Liao et al. (2002) tested several clustering algorithms including K-means, fuzzy c-means and genetic clustering to multivariate simulation time series data of unequal length with the objective to form a discrete number of states. The original time series data were not evenly sampled and made uniform by using the simple linear interpolation method.

Van Wijk and Van Selow (1999) performed an agglomerative hierarchical clustering of daily power consumption data based on the root mean square distance. The patterns of clusters distribution were explored with calendar-based visualisation. Kumar et al. (2002) evaluated a distance function that is derived from the assumed independent Gaussian models of data errors. They use the hierarchical clustering method to group seasonality sequences and related sequences into a desirable number of clusters. The experimental results with simulated data and retail data indicated that the new method performs better than both k-means and Ward's method which do not incorporate data errors in terms of (arithmetic) average estimation error in the formulation and calculation of the model.

However the model assumed that data used will have been pre-processed to remove the effects of non-seasonal factors and normalised to enable comparison of sales of different items on the same scale.

In the area of DNA, Möller-Levet et al. (2003) derived the STS distance as a measure of the similarity in shape formed by the relative change in the level of amplitude and the corresponding temporal information of uneven sampling intervals. All series are considered sampled at the same time points with corresponding time stamp. Through the incorporation of STS distance into the standard fuzzy c-means algorithm, they improved the equations for computing the membership matrix and the prototypes which evolved into a fuzzy time series clustering algorithm.

Liao (2005) developed a two-step procedure for clustering multivariate time series of equal or unequal length. The first part applies the k-means or fuzzy c-means clustering algorithm to time independent data in order to transform multivariate real-valued time series into univariate discrete-valued time series. The converted variable is used in the model and interpreted as a state variable process. The second part employs the k-means or FCM algorithm again to group the converted univariate time series, with the inputs as transition probability matrices, into a number of clusters. The traditional Euclidean distance is used in the first step, while various other distance measures such as the symmetric version of Kullback-Liebler distance can be employed in the second step.

Clustering based on raw data implies working with high-dimensional space – especially for data collected at fast sampling rates. It is also not desirable to work directly with the raw data that are highly noisy. Several feature-based clustering methods have been proposed to address these concerns. Though most feature extraction methods are generic in nature, the extracted features are usually application dependent. That is, one set of features that work well on one application might not be relevant to another. Some studies even take another feature selection step to further reduce the number of feature dimensions after feature extraction.

Shaw and King (1992) clustered time series indirectly through the application of two hierarchical clustering algorithms, the Ward's minimum variance algorithm and the single linkage algorithm, to normalised spectra. The spectra were the calculated and derived from the original time series with the means adjusted to zero. The principal component analysis is used to filter spectra which were also clustered. The conclusion was found that the use of 14 most significant eigenvectors could achieve comparable results. The Euclidean distance was used.

For clustering using a set of dynamic structures which belong specifically to the class of ARIMA invertible models, Piccolo (1990) introduced the Euclidean distance between the corresponding autoregressive expansions as the metric. This metric satisfies the classical properties of a distance such as non-negativity, symmetry and triangularity. He further discussed six additional properties of the metric. The distance matrix between pairs of time series models will then be processed by a complete linkage clustering method to construct the dendrogram.

Baragona (2001) evaluated three meta-heuristic methods for separating a set of time series into clusters in a manner that maximises the cross-correlation maximum absolute value between each pair of time series that belong to the same cluster is greater than some minimum threshold and the k-min cluster criterion is minimised with regards to a specified number of clusters.

The cross-correlations are computed from the residuals of the models of the original time series. After evaluating all the various methods, tabu search was discovered to

perform better than single linkage, pure random search, simulation annealing and genetic algorithms given results based on a simulation experiment on ten sets of artificial time series created using low-order univariate and vector ARMA models.

Kalpakis et al. (2001) examined the clustering of ARIMA time-series, through the Euclidean distance between the linear predictive coding (LPC) cepstral of two time-series as the key dissimilarity measure. The cepstral coefficients for an AR(p) time series are then calculated using the auto-regression coefficients. The partition uses medoids method and the k-medoids algorithm was chosen with the clustering outcome evaluated using the cluster similarity measure and Silhouette width. Based on a test of four datasets, they demonstrated the superiority of LPC cepstrum which provides higher discriminatory power to differentiate one time series from another compared to other widely used methods such as the Euclidean distance between the DFT, DWT, PCA and DFT of the auto-correlation function of two time series.

Every clustering algorithm relies on a suitable measure to compute either distance or similarity between two time series. Certain particular measure might be more appropriate than another depending on the type of time series in question. Most clustering algorithms are iterative in nature and rely on a suitable criterion to determine whether clustering obtained is in good condition to stop the iterative process. There were extensive discussions about the problems of using the Euclidean distance in the comparison of time series (Lin et al., 2003). However, most of the criticisms are targeted at the pattern recognition of time series behaviours. Subsequent research indicated that in certain cases where pure pattern recognition is not the main objective, direct application of the Euclidean distance measure as the main clustering input does not affect the effectiveness of the technique (Chen, 2005).

The hierarchical clustering method works by grouping data objects into a tree of clusters. There are two types of hierarchical clustering methods. The two major clustering processes are agglomerative or divisive depending upon whether the bottom-up or top-down strategy is followed. The agglomerative approach is more commonly used than the divisive method. The algorithm starts by having each object in its own cluster and start merging the atomic clusters into larger and larger clusters until all the objects are in a single cluster. The single linkage approach measures the similarity between two clusters using the similarity of the closest pair of data points between the clusters. The closest two clusters are merged and the process repeats merging until all the clusters forms one cluster. The Ward's minimum variance approach differs by merging the two clusters that minimally increase the value of the sum-of-squares variance. At every step, all possible mergers of two clusters are tried and the one with the smallest increase is selected. The agglomerative hierarchical clustering method often suffers from adjustment problem. Hierarchical clustering is not restricted to cluster time series with equal length and can be extended to series of unequal length with the appropriate distance measure.

3 IBOR rate mechanism

The IBOR rate mechanism is extremely similar across the various countries and jurisdictions. The most common cited IBOR is the LIBOR which was standardised in 1984 by the British Bankers' Association (BBA) as the main reference rate for numerous securities which includes syndicated loans, futures contracts and forward rate agreements. The LIBOR is also used to as the reference for unsecured instruments between banks in

London and elsewhere globally. Globally, all the IBORs behave similarly to LIBOR in that they are quoted daily for several major currencies.

The various IBORs rely on a panel of selected banks to provide daily rate quotes for the calculation of the IBOR. The banks are selected based on a variety of criteria such as size of operation, reputation as well as capabilities and knowledge of the currency concerned. Typically, the biggest banks operating in the particular currency will be consulted for the rates.

Currently, LIBOR is defined as the rate at which an individual contributor panel bank could borrow funds, were it to do so by asking for and then accepting inter-bank offers in reasonable market size, just prior to 11:00 London time (BBA, 2012).

This definition is further broken down into the following sections (BBA, 2012):

- the submitted rate must be formed from that bank's estimated cost of funds in the interbank market
- contributions must represent rates formed in London only
- contributions addresses only the currency concerned and does not seek to address the cost of producing one currency by borrowing in another currency and accessing the required currency via the foreign exchange markets
- the rates must be submitted by members of staff at a bank with charged with the management of a bank's cash
- the 'funds' is defined as the unsecured interbank cash or cash raised via the issuance of interbank certificates of deposit.

Every contributor panel bank is required to directly input its data no later than a given time for each day that the capital market is open. After the given time, an appointed agent will then process the calculation. Usually, for each maturity, the agent will eliminate the highest and lowest X% of all the quotes collected to eliminate the outliers. The remaining rates will be averaged and rounded to three decimal places. It is precisely this approach that resulted in the 2012 LIBOR scandal.

4 LIBOR scandal

The LIBOR scandal is a series of events that consist of fraudulent actions by the bank with regards to their behaviours with regards to the LIBOR. The incidents arose when the banks falsely inflate or deflate their rates so as to profit from trades and implant false impression of creditworthiness.

As mentioned in the earlier section, the banks are expected to submit the actual interest rates they are paying (or based on their expectation) for borrowing from other banks. The LIBOR behaves like an overall assessment of the health of the financial system and acts as the litmus test for financial health of the banks. If the banks being polled feel confident about the state of affair, they will report a low number and vice versa in a situation of low confidence.

As LIBOR is used in US derivatives markets (and elsewhere) as a benchmark, any attempt to manipulate LIBOR is considered to be an attempt to manipulate US derivatives markets and violates the US laws. Given that many financial products rely on LIBOR as the reference rate, any manipulation of the submissions for the calculation of

the rates can have strong and significant consequent negative impacts on consumers and financial markets worldwide.

There were strong debates about how the submissions could have affected the LIBOR. Lively discussions erupted online and two major camps emerged. The first group of thinkers argued that the nature of the LIBOR calculation makes it impossible for any one single bank to manipulate it and any manipulations must be the product of several banks colluding with one another (Persaud, 2012). The second group of thinkers believes that the LIBOR has been manipulated and that any bank can manipulate the rates (Smith, 2012).

The discussions raised a number of key issues for attention. In the following section, we will be addressing several of these important and non-trivial questions. Specifically, the questions to answer are as follows:

1 Can LIBOR be manipulated by one single bank?

2 Were the banks involved in LIBOR colluding? If so, who among the banks and what are some tell-tale signs?

3 How similar is SIBOR to LIBOR and was it manipulated like LIBOR?

5 Experiment

To detect irregularities in the IBOR rates, we will first extract data from suitable sources. To achieve this, the authors have extracted the LIBOR and SIBOR submission data from Bloomberg data services. To cover the period when there were data manipulation, data for LIBOR and SIBOR for one month maturity were collected for the period 2005 to 2012. This is to ensure consistencies in comparison.

To answer whether LIBOR can be manipulated by a single bank, we have to first understand the mechanism of the LIBOR calculation which is based on the concept of trimmed mean. The trimmed mean is a type of statistical measure of central tendency which is similar to the mean and median. It is calculated by removing both ends of the extreme values. Let us look at the following example.

Table 1 Theoretical LIBOR scenario (no manipulations) (see online version for colours)

Bank	*Rates*
1	3.0026
2	3.0106
3	3.0235
4	3.0312
5	3.0358
6	3.0434
7	3.0562
8	3.0601
9	3.0658
10	3.0961
LIBOR	3.04168

Assuming that we have ten banks and under normal circumstances, we have the rate submitted in Table 1.

Let us assumes that Bank 9 wishes to lower LIBOR and submits a very low rate. Table 2 is the scenario.

Table 2 Theoretical LIBOR scenario (with manipulations) (see online version for colours)

Bank	*Rates*
1	3.0026
2	3.0106
3	3.0235
4	3.0312
5	3.0358
6	3.0434
7	3.0562
8	3.0601
9	3.0000
10	3.0961
LIBOR	3.0334

From the example, we can see that the LIBOR has been successfully been lowered. Thus the answers to questions 1 and 2 are clear, any single bank can through its submission of rates manipulate the rates without having to even collude with any other banks. Thus we can safely conclude that the existing approach can be vulnerable to manipulation.

To prevent any manipulations of the IBORs, the only possible measure is to attempt to detect anomalous behaviour in the rate submissions. This is considered simple to achieve by conducting time series clustering on the data. This is because any banks attempting to manipulate the rates will display behaviours which are distinctively different from the rest of banks. Early splits in the dendrogram of cluster analysis will reveal unique clusters of banks which behave very differently from others. This can be seen from the results in the next section.

6 Results

We first begin our analysis on the 2005 and 2008 data where there were previous studies that Barclays had actually attempted to manipulate the rates. In Figure 1 are the results of the cluster analysis performed on the 2005 LIBOR data.

From the figure, we can observe that there are several distinct groups. However, Barclays does not display any significant deviation from the rest of the banks for 2005 for one month maturity. The most distinctive groups are the BTMU/NORIN, BOA and DB/RBS. From the distance measured, it can be seen that most of the rates are quite similar to one another. However, the picture could not have been more distinct in the year of 2008 as shown.

Figure 1 LIBOR 2005 cluster analysis

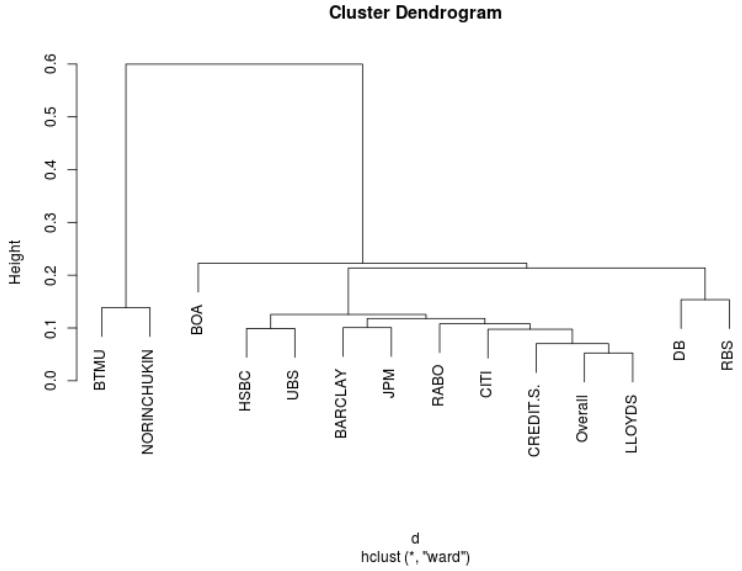
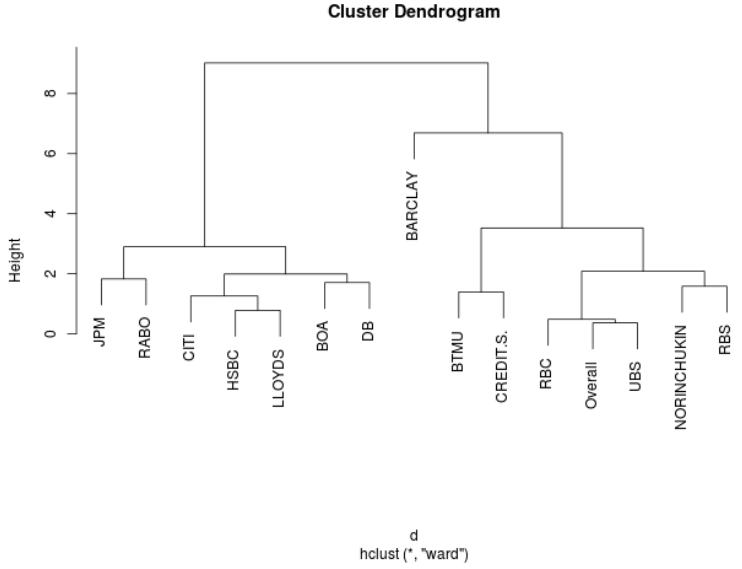


Figure 2 LIBOR 2008 cluster analysis



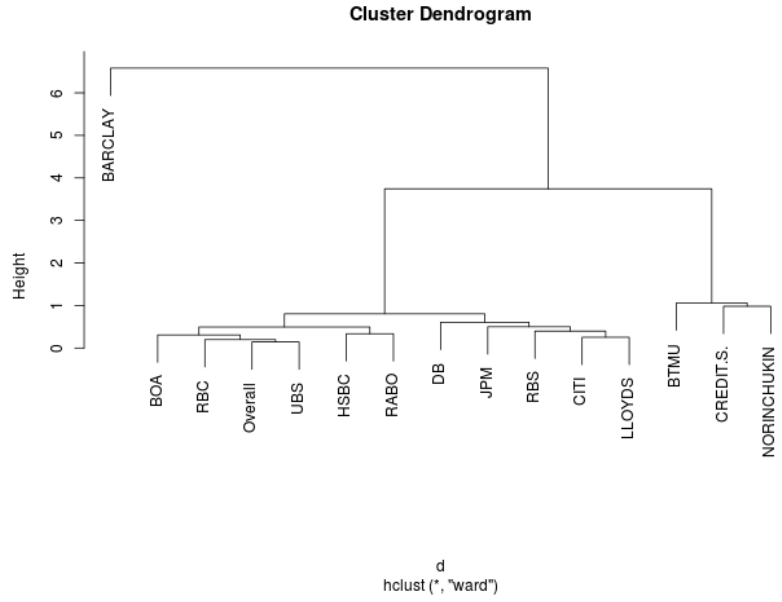
From the analysis, we can observe that Barclays display significant deviations from the rest of the banks for 2008 for one month maturity. While there are two major groups of banks, they are distinctively different from one another and can be attributed to the financial distress from the crisis. Banks such as Citibank and JP Morgan were under less severe market conditions as compared to Credit Suisse as they were assisted by the TARP. However, we can see that Barclays is definitely different from the rest of the

banks as shown in Table 3. Thus, for question 2, we can conclude that there was no obvious collusion between the banks.

Table 3 LIBOR rates 2008 (average daily) (see online version for colours)

Banks	*Rates*
JPM	2.600
RABO	2.607
DB	2.616
CITI	2.643
BOA	2.648
LLOYDS	2.653
HSBC	2.656
UBS	2.673
RBC	2.673
Overall	2.674
RBS	2.686
CREDIT S.	2.727
NORINCHUKIN	2.727
BTMU	2.738
BARCLAY	2.783

Figure 3 LIBOR 2008 quarter three cluster analysis



The cluster analysis result is also verified by examining the data from 2008 quarter 3 where Barclays has behaved differently from the rest of the banks. We can see that almost all the banks are similar compared to Barclays indicating that there were certain

actions which were anomalous. Let us review the SIBOR for 2007 and 2008 financial crisis period.

Figure 4 SIBOR 2007 cluster analysis

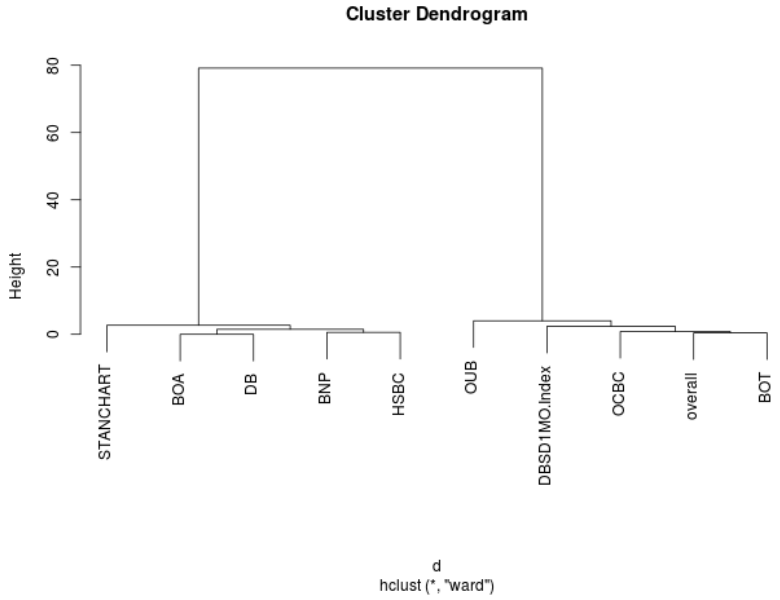
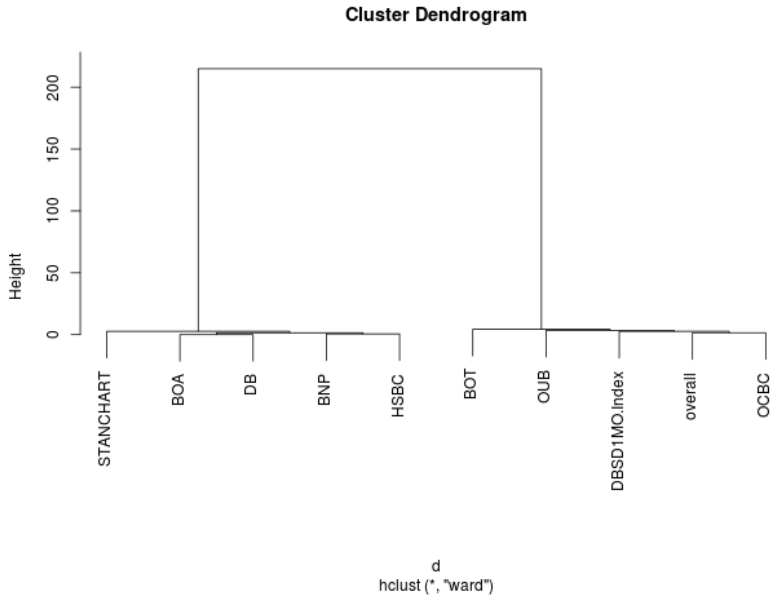


Figure 5 SIBOR 2008 cluster analysis



The information shown in SIBOR charts (Figures 4 and 5) does not reveal any distinct individual which is very different from the rest of the banks. There are two distinct

groups though. The first group comprises of the major local banks while the second group consists of overseas or international banks.

Table 4 SIBOR rates 2007 (average daily) (see online version for colours)

Banks	*Rates*
DBSD1MO index	2.627
OUB	2.648
Overall	2.658
OCBC	2.662
BOT	3.375
BNP	3.438
HSBC	3.469
BOA	3.500
DB	3.500

Table 5 SIBOR rates 2008 (average daily) (see online version for colours)

Banks	*Rates*
DBSD1MO index	1.070
OUB	1.130
Overall	1.152
OCBC	1.152
BOT	1.211
STANCHART	3.375
BNP	3.438
HSBC	3.469
BOA	3.500
DB	3.500

The major reason for the differences is the financial distress experienced by the international banks in their home country lending markets. Another reason is the lack of exposure of the local banks to the toxic portfolio components of subprime lending. There is no evidence to suggest that they were attempting to manipulate the rates by colluding with one another. Thus, for question 3, we can safely say that SIBOR is not manipulated during the financial crisis.

7 Conclusions

In the above exercise, the cluster analysis technique was able to identify Barclays as the bank behaving anomalously from the rest of the banks. The dendrogram identifies the

most obvious bank which is not behaving according to market norm. In the LIBOR case for 2005 and 2007, there were no obvious collusions with the exception of Barclays. SIBOR was not affected by rate manipulation between 2007 and 2008. However, the technique has some limitations and weaknesses as well.

The cluster analysis while being able to detect rogue behaviours is not able to uncover any collusions that are extremely well planned and or affect the rates mildly. As demonstrated in the sections above, any bank can influence the rates. Consider the case where several banks wish to influence the rate by 0.01%, that can be done easily by quoting the same lowest rate together and thus forming a group. In the dendrogram, they will be found as a group. The group will be so big that they will be similar to the SIBOR case. Although further investigation into the manner of grouping might reveal anomalies, it cannot be directly inferred from the diagram nor through any clustering techniques. Another problem is the detection of a false positive case. It is possible for a bank to be abnormally low or high in certain circumstance which may result in a false positive detection. The other more important problem is the behaviour of the manipulation tactics in short period of time. Unless there are obvious and extended period of manipulation, it is unlikely that any statistical tests or techniques will be able to discover any such issues. The time period in question is also important. The results from monthly to weekly as compared to quarterly will differ drastically and the suitable time frame selected for analysis is critical to ensuring that the banks has manipulated for a period of time before they can be viewed as systematically manipulating the IBORs. In the SIBOR case, the grouping is easily understood as it reflects the origin of the banks. As such, we would not call for additional investigation.

In June 2013, the Monetary Authority of Singapore (MAS) disciplined 20 banks and revealed that 133 traders tried to manipulate three interest rate and foreign exchange benchmarks. The MAS probe of the Singapore interbank lending rate was sparked by the revelation that banks sought to manipulate LIBOR (*Financial Times*, 2013). It was reported that the authorities found improper traders' behaviours that reflected a lack of professional ethics. However, there was no conclusive finding that SIBOR, swap offer rate (SOR) and FX benchmarks were successfully manipulated. The results of our study would lend support to the findings of the authorities. Thus we can conclude that the time series clustering method did have its advantages in discovering any blatant attempts to influence the IBORs.

Further research will be needed to enhance the analysis so that the display of collusion can be incorporated into the analysis. At this point of time, the model does not attempt to incorporate or measure the level of collusion between the banks. This might be useful if we can do this as it will reflect the lack of free market and lead to further studies on the market behaviours when we have trouble. Further investigation will also be needed to understand how many banks colluding together will nullify the technique as well.

Acknowledgements

The authors would like to acknowledge the efforts of the reviewers and editor during the review process and the suggestions provided by the reviewers.

References

Baragona, R. (2001) 'A simulation study on clustering time series with metaheuristic methods', *Quaderni di Statistica*, Vol. 3, pp.1–26.

BBA (2012) *Guide to LIBOR*.

Chen, J.R. (2005) 'Making subsequence time series clustering meaningful', *Fifth IEEE International Conference on Data Mining*, IEEE, November, p.8.

Financial Times (2013) [online] http://www.ft.com/intl/cms/s/0/fed38a0a-d4d5-11e2-b4d7-00144feab7de.html (accessed 16 April 2013).

Kalpakis, K., Gada, D. and Puttagunta, V. (2001) 'Distance measures for effective clustering of ARIMA time-series', *Proceedings of the IEEE International Conference on Data Mining, 2001, ICDM 2001*, IEEE, pp.273–280

Košmelj, K. and Batagelj, V. (1990) Cross-sectional approach for clustering time varying data. *Journal of Classification*, Vol. 7, No. 1, pp.99–109.

Kumar, M., Patel, N.R. and Woo, J. (2002) 'Clustering seasonality patterns in the presence of errors', *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July, ACM, pp.557–563.

Liao, T.W., Bolt, B., Forester, J., Hailman, E., Hansen, C., Kaste, R.C. and O'May, J. (2002) 'Understanding and projecting the battle state', in *23rd Army Science Conference*, December, pp.2–5.

Liao, W.T. (2005) 'Clustering of time series data–a survey', *Pattern Recognition*, Vol. 38, No. 11, pp.1857–1874.

Lin, J., Keogh, E. and Truppel, W. (2003) 'Clustering of streaming time series is meaningless', *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ACM.

Möller-Levet, C.S., Klawonn, F., Cho, K.H. and Wolkenhauer, O. (2003) 'Fuzzy clustering of short time-series and unevenly distributed sampling points', *Advances in Intelligent Data Analysis V*, pp.330–340, Springer, Berlin, Heidelberg.

Persaud, A. (2012) [online] http://www.voxeu.org/article/notes-scandal-libor (accessed 3 December 2012).

Piccolo, D. (1990) 'A distance measure for classifying ARIMA models', *Journal of Time Series Analysis*, Vol. 11, No. 2, pp.153–164.

Shaw, C.T. and King, G.P. (1992) 'Using cluster analysis to classify time series', *Physica D: Nonlinear Phenomena*, Vol. 58, No. 1, pp.288–298.

Smith, Y. (2012) [online] http://www.nakedcapitalism.com/2012/07/libor-scandal-apologist-avinash-persaud-displays-inability-to-do-math.html (accessed 12 January 2013).

Snider, C. and Youle, T. (2010) *Does the LIBOR Reflect Banks' Borrowing Costs?*, SSRN 1569603.

Van Wijk, J.J. and Van Selow, E.R. (1999) 'Cluster and calendar based visualization of time series data', *Proceedings on the 1999 IEEE Symposium on Information Visualization, 1999 (Info Vis' 99)*, IEEE, pp.4–9.

Real Time Process Compliance Checking using Nomenclature Approach

Murphy Choy, Ma Nang Laik and Koo Ping Shung
School of Information System
Singapore Management University
80 Stamford Road
Singapore 178902

Abstract

Process compliance has experienced rapid development in the wake of several serial scandals in both financial institutions such as RBS and Société Générale as well as non-financial institution such as Enron and Olympus. These scandals resulted in legislation enacted specifically to address these issues and prompted the businesses to develop more systematic approaches to the design of business processes which adhere to the reality as well as satisfying the compliance requirements. There are several researches into the active use of logical search approaches in checking whether the process meets the requirements. In general, most of the research focuses on the extension of the logical pattern recognition. However, most of the compliance research focuses only on the examination of business process flow during the design phase and not operation. In this paper, the authors will like to propose the novel use of Nomenclature approach to evaluate business processes as they occur. We also illustrate using a real-world case study on how this new approach also facilitate the checking of compliance in real time.

Keywords
Compliance, Process Analysis, Deontic Logic, Nomenclature, Business Process Extraction

Introduction

Modern businesses face a huge number of challenges to maintain their competitive edge. Companies must achieve the expectations set by the share-holders and remain profitable while maintaining positive customer relations. With globalization and growth of digit medias, companies are confronted with dynamic competition conditions. This resulted in large investments in information technology which is necessary to stay competitive and remain in business. To further enhance their competitiveness, many enterprises have demonstrated interest in business process management. Most businesses wished to model, automate, optimize and monitor their businesses processes to improve their customer satisfaction, increase revenue or reduce the operational inefficiency. In the recent years, there have been an increased acceptance and adoption of business process management systems.

The development has been further fuelled by the growing number of regulatory requirements imposed on business operations as a result of corporate scandals. The most prominent legislations are the Gramm-Leach-Bliley Act (US, 1999) and the Sarbanes Oxley Act (SOX) (US, 2002). While these regulations have coverage across industries, there are also industry-specific regulations such as the Basel III accord (Basel, 2011) or the European Money Laundering Regulation (UK, 2003). Demonstration of compliance with specific legal legislations and international standards requires the company to document their existing business processes as well as verifying that the processes conform to legislation. Most enterprises regard such documentation requirements and strict adherence to be a business cost, they also recognize the opportunity to identify and document their formal and informal processes to render their executions more efficiently and effectively. For huge enterprises with several divisions and thousands of different business processes, this is a major challenge both in documentation as well as process monitoring.

Most enterprises that operate in heavily regulated industries, such as financial services or health care, are controlled by a huge number of regulatory requirements that define most of their operations. As these implemented requirements needed to be enforced by a multitude of internal business and IT controls, several regulations recommend the use of industry standards, such as COBIT (Control Objectives for Information and Related Technologies) (COBIT, 2005) and ITIL (Information Technology Information Library) (ITIL, 2006), in the implementation of any enterprise IT system. These standards consist of well-defined abstract process definitions which can be tailored according to individual needs.

With increasing number of legislations and standards, companies needed a comprehensive compliance-management approach to manage the design of new processes as well as ensuring the conformance of process. The need to be able to understand the impact of new regulations for their business and processes becomes ever more important. As business processes are increasingly managed using Business Process Management Systems (BPMS), any regulatory requirements that require changes to the structure of any workflows will directly impact business process modeling.

Whenever there is new regulatory requirement, a company will need to understand the impact and react accordingly. There are three possible outcomes,

1. Adaptation or removal of business processes.
2. Creation of new business processes.
3. No change is needed.

Business processes that are automated through BPMS can implement IT processes and controls, as stipulated by ITIL or COBIT and address existing legislation requirements. However new regulatory requirements cannot be readily or easily assessed with these frameworks. For large companies with thousands of business processes deployed on the BPMS, the ability to quickly assess the compliance of existing process definitions is extremely important. In this paper, we describe a nomenclature approach that caters to static verification of business process models against a predefined set of regulatory requirements such as execution order of certain tasks. The nomenclature approach also performs historical process checking as well as real time checking. Our approach will assist the business to identify non-compliant business processes before implementation and identify non-compliance business process during operations.

Importance of automated verification of business processes

Before compliance can be enforced, the effect of every new legislation on existing business processes needs to be identified. Our approach identifies the non-compliance processes at the relevant stages which provides a useful tool to ensure that new requirements are adapted into the processes. Hence, the contribution of our method is three folds:

1. With automated verification of business processes during design phase and operation phase, our approach increases efficiency of detection of non-compliant processes and lower the risk of the event of non-compliance. (Abrahms et. al., 2007; Giblin et. al., 2005; Ly et. al., 2011; Governatori and Rotolo, 2010)
2. Through automation of a manually tedious task, our method also reduces the cost of inspecting business process models for compliance especially during operation.
3. The real time updating of the process compliance allows management to have a better and more in-depth understanding of the problems that the business faces.

Nomenclature Approach

Nomenclature is defined as a systematic approach to the naming of items in the area of science or arts by individuals or community. It can also refer to the systematic naming of items according to taxonomy. In most scientific disciplines, there are established standards of nomenclature. Nomenclatures are used predominantly in Mathematics, Biology as well as Chemistry. The use of nomenclatures in taxonomy allows for easy tracing of the origin of the species. Nomenclatures also allow for easy identification of similar components between two chemicals or species.

In business process management, most of the business processes were mapped out in BPMS software in a logical decision flow manner such as the use of BPELs (Business Process Extraction Language) (Antoniou et. al., 2005; Ghose and Koliadis, 2007), BPCL (Business Process Compliance Language) or BPSL (Business Property Specification Language) . These models are extremely useful in understanding whether the flow is sensible and what are the components of the process. Other software uses symbolic logic such as Standard Deontic Logic (Alberti, 2004; Alberti, 2005) to assist in the checking of the process.

In this paper, we will discuss about the modification of a process flow with the addition of elements of nomenclature approaches from other area of sciences. Below we have several simple processes,

$$a_1 \Rightarrow a_2$$

$$a_1 \Rightarrow a_2 \Rightarrow a_1 \Rightarrow a_2 \Rightarrow a_1 \Rightarrow a_2$$

$$a_1 \Rightarrow a_2 \Rightarrow a_1 \Rightarrow a_2 \Rightarrow a_1 \Rightarrow a_2 \Rightarrow a_1 \Rightarrow a_2$$

where both a_1 and a_2 are two separate processes which are linked to one another and they can be repeated sequentially. In most cases, these processes will be viewed as separate processes. However, this is inappropriate as they are essentially the same process repeated multiple number of times. Under normal circumstances, the common compliance software will have to validate three separate processes. By simplifying these processes which are repeated for multiple times, we can express them in summarized form below,

$$[a_1 \Rightarrow a_2][1]$$

$$[a_1 \Rightarrow a_2][3]$$

$$[a_1 \Rightarrow a_2][4]$$

The process can be generalized in the form below which reduces the number of processes that needs to be checked to one.

$$[a_1 \Rightarrow a_2][m] \text{ Where} m \in \mathbb{Z} \text{and } m \geqslant 1$$

While we have discussed about the case for multiple events processes which are repetitive, the case for single event repetition has not been discussed. To enable one to distinguish a single event from repeated process, we will be using the notation as below.

$$\dots a[n] \dots$$

Algorithm for extraction and simplification

As described in the prior section, the search for repeated process and events needs to follow a simple logic to simplify the process. Below is the pseudo-code for algoithm for searching the process flow as well as simplifying the process.

Let us first assume the following general process

$$a_1 \Rightarrow a_2 \dots a_{k-1} \Rightarrow a_k$$

Given the process has k events, the maximum length for repeatable substring within a longer string will be $\frac{(k-1)}{2}$ if k is an odd number and $\frac{k}{2}$ when k is an even number (Lucian and Smyth, 2011), if the maximum theoretical possible length is not given by the process owner. Once the length is established, we will then initiate the search from this length to the base minimum length of 1 which is similar to the behaviour of suffix tree (Ukkonen, 1995). For each level of process search, we will also create a holder to indicate the start of the repeated process. This level of process will then set all lower levels to 0 so as to prevent multiple level simplifications. At the same time, we have to set the constraint where all the events in the process are not the same.

Translating the above information, the pseudo code form is as follow.

Let X_k be the event X at position k in a chain of events.
Let i be the number of events for a particular process.
Let j be the number of levels for a particular process (maximum length for repeatable subsring).
Let L_{kj} be the repeat flag at position k for level j in a chain of events.
Let h be the number of repetitions of a sub-process.

1. Set Max Length (j) as $\frac{(i-1)}{2}$ if j is odd or $\frac{(i)}{2}$ if j is even, otherwise j as given.
2. Create j levels of holder with length i.
    a. For level j till 1,
        i. For position k from 1 to i-j,
        ii. Ensure $X_k = X_{k+j}$
        iii. Repeat the check till $X_k = X_{k+j-1}$
        iv. If j > 1 then check for position k, $X_k \neq X_{k+1}$
        v. If so, $L_{k,j} = 1$ else $L_{k,j} = 0$
3. For level j till 1,
    a. For position k from 1 to n-m+1,
        i. If $L_{k,j} = 1$ and $L_{k,j-1} = 1$, then $L_{k,j-1} = 0$
4. For level j till 1,
    a. For position k from 1 to n-m,
        i. Check whether for $L_{k,j} = 1$
        ii. If yes then do the following
            1. Repeat the check on $L_{k+j*h,j}$ until $L_{k+j*h,j} = 0$ where h indicates the number of repetitions
            2. Condense the process from position k to k+j*h with the repetitions denoted as h+1.
            3. Move the search to k + j*h +1 position.
5. The condensed process is then created from the iteration process.

Let us look at an example where i = 8 and j = (8)/2 = 4.

$P01 \Rightarrow P02 \Rightarrow P01 \Rightarrow P02 \Rightarrow P03 \Rightarrow P05 \Rightarrow P06 \Rightarrow P06!$

Let us assume that the maximum length of any single process is 2. From the algorithm, we can evaluate the process and notice that components 1, 2 and 3 are the same. This reduces the level 2 process to this.

$$(L_{2,1} = 1) \Rightarrow (L_{2,2} = 1) \Rightarrow (L_{2,3} = 0) \Rightarrow (L_{2,4} = 0) \Rightarrow (L_{2,5} = 0) \Rightarrow (L_{2,6} = 0) \Rightarrow (L_{2,7} = 0) \Rightarrow (L_{2,8} = 0)$$

Repeating the process for level 1, from the algorithm, we can evaluate the process and notice that components 7 and 8 are the same. This reduces the level 1 process to this.

$$(L_{1,1} = 0) \Rightarrow (L_{1,2} = 0) \Rightarrow (L_{1,3} = 0) \Rightarrow (L_{1,4} = 0) \Rightarrow (L_{1,5} = 0) \Rightarrow (L_{1,6} = 0) \Rightarrow (L_{1,7} = 1) \Rightarrow (L_{1,8} = 0)$$

Once we have identified the levels, we can now proceed to do the clean up.

$$(L_{2,1} = 1) \Rightarrow (L_{2,2} = 1) \Rightarrow (L_{2,3} = 0) \Rightarrow (L_{2,4} = 0) \Rightarrow (L_{2,5} = 0) \Rightarrow (L_{2,6} = 0) \Rightarrow (L_{2,7} = 0) \Rightarrow (L_{2,8} = 0)$$

$$(L_{1,1} = 0) \Rightarrow (L_{1,2} = 0) \Rightarrow (L_{1,3} = 0) \Rightarrow (L_{1,4} = 0) \Rightarrow (L_{1,5} = 0) \Rightarrow (L_{1,6} = 0) \Rightarrow (L_{1,7} = 1) \Rightarrow (L_{1,8} = 0)$$

From the clean up processes, we start from level 2 repeated process and start condensing the process.

$[P01 \Rightarrow P02][2] \Rightarrow P03 \Rightarrow P05 \Rightarrow P06 \Rightarrow P06$

From the clean up processes, we start from level 1 repeated process and start condensing the process.

$[P01 \Rightarrow P02][2] \Rightarrow P03 \Rightarrow P05 \Rightarrow P06[2]$

In the next section, we will discuss about the application of this approach to a logistic company and how it assisted them in the discovery of processes which are non-compliant.

Case Study of Logistic Company

In the case of the logistic company A, they have several delivery services with many business processes controlling the operation of the company. Because of the nature of the business, certain legislative requirements force the company formulate their existing business process around those legislation. At the same time, because of the nature of the service rendered, there are contractual agreement on the service level making it important to analyze the business processes for any anomalous behaviours.

Currently, the business process is mapped out using the BPMS software and most of the booking of services are done electronically. The operation uses electronic devices to scan the delivery process and track the movement of items. The devices capture all the process status and update them to the system. However, there are no visibility to the compliance level of the business processes. The existing BPMS does not allow the company to check their existing processes.

Using the new nomenclature approach, we have mapped out the existing processes and converted them into general processes. To account for several possible repeated processes and sub-processes, we have generated several iteration of the different processes. Using the nomenclature forms, we then analyze their existing structures. Without using the generalized form, there are almost 30,222 different processes which needed to be mapped to the BPMS systems. This is tremendous amount of work and passing them through the system individually to test the logic flow is also a computational intensive work. Using the generalized nomenclature approach, we reduce the processes to 10,272 which

is around 65% reduction in the number of processes. To test the efficiency of the algorithms, we conducted a test of the run time of the algorithms for 6 business areas and estimate the efficiency improvement from deontic to nomenclature approach.

To evaluate the runtime efficiency of the new approach, we conducted simulation run of the algorithms and compared the runtime efficiency using paired sample t-test. The simulation first virtualizes two separate compliance checking systems on two separate machines. Both systems are linked to a central database that asynchronously passes the data to the compliance checking systems. The data base is fed with real life information at regular interval which simulates the real operations of such systems. The run time is measured by the amount of time needed to process the data to check for compliance after each update. The simulation is ran 100 times which simulate 100 days of operation for both systems and the paired sample t-test used to compare the differences in the runtime. Six business areas were selected which represent 90% of all the transactions. The six business areas are same day parcel, same day mail, normal mail, express mail, bulk parcel and letter deliveries.

The paired difference test is a type of location test used when comparing two sets of measurements to assess whether their population means differ. A paired difference test uses additional information about the sample that is not present in

an ordinary unpaired testing situation, either to increase the statistical power, or to reduce the effects of confounders. Specific methods for carrying out paired difference tests are, for normally distributed differences the paired t-test where the population standard deviation of difference is not known.

Let R_i be the run time for deontic approach run i.
Let W_i be the run time for nomenclature approach run i.
Let μ_0 be the population mean to be tested against.
Let $\bar{X}_D$ be the sample difference mean.
Let S_D is the sample standard deviation of the sample
Let n be the sample size.

The t-test will test the null hypothesis (H_0) which assumes that population mean is equal to a specified value μ_0. The alternate hypothesis (H_1) assumes that population mean is not equal to a specified value μ_0.

$$H_0: \bar{X}_D = \mu_0$$
$$H_1: \bar{X}_D \neq \mu_0$$

Let X_i be the run time difference between deontic and nomenclature approach for run i.

$$X_i = R_i - W_i$$

Below is the calculation for $\bar{X}_D$.

$$\bar{X}_D = \frac{\sum_{i=1}^{n}(X_i)}{n}$$

Below is the calculation for S_D.

$$S_D = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X}_D)^2}{n}}$$

Below is the paired samples t-test statistics.

$$t = \frac{\bar{X}_D - \mu_0}{\frac{S_D}{\sqrt{(n)}}}$$

If the test statistics is significant and $\bar{X}_D$ is greater than 0, then nomenclature approach is superior. Otherwise the test statistics is significant and $\bar{X}_D$ is lesser than 0, then deontic approach is superior. Below are the various test results.

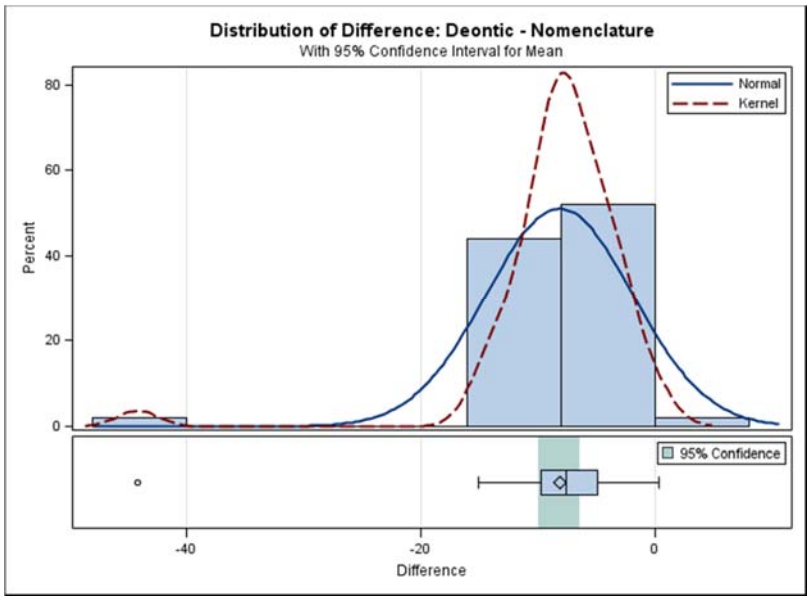


Figure 1: T-Test Results between Deontic and Nomenclature approaches for same day delivery

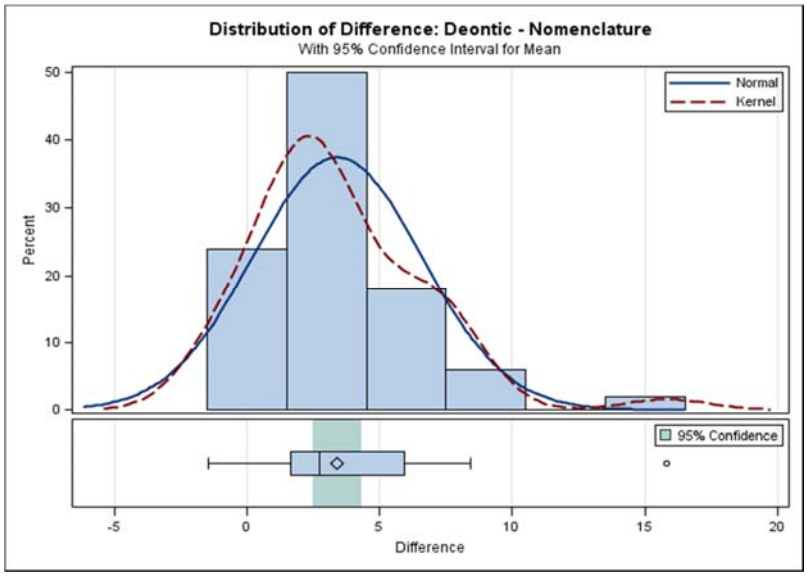


Figure 2: T-Test Results between Deontic and Nomenclature approaches for same day mail

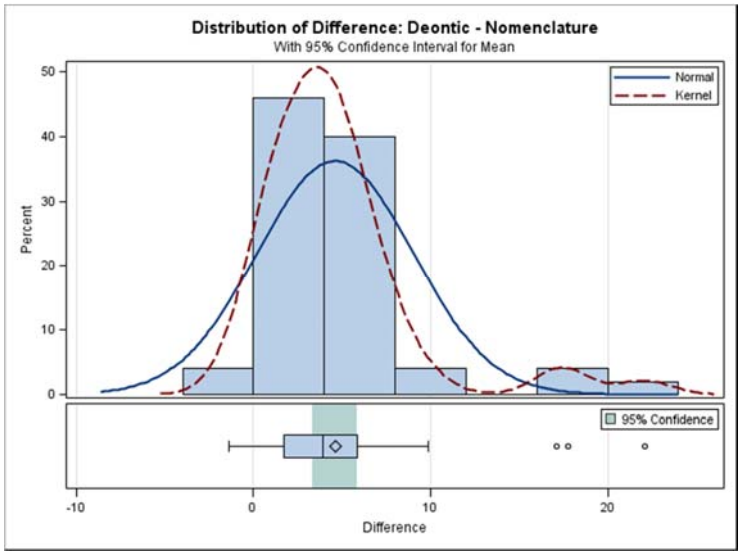


Figure 3: T-Test Results between Deontic and Nomenclature approaches for normal mail

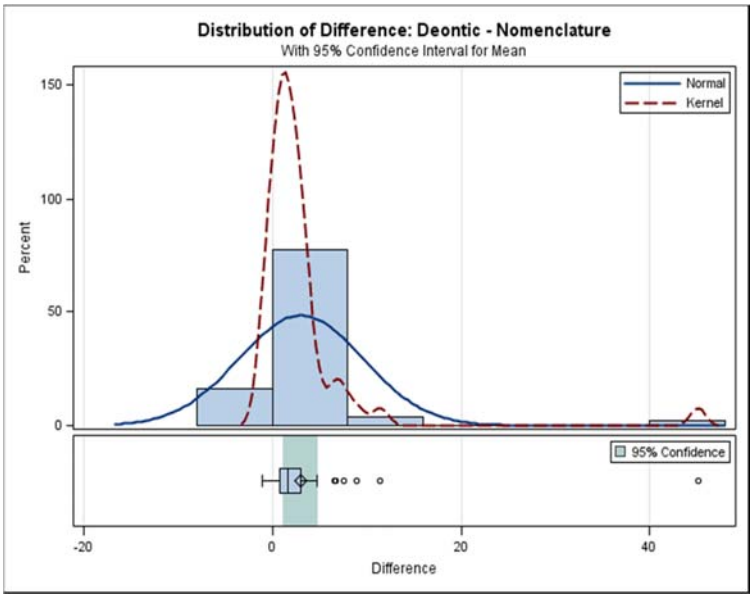


Figure 4: T-Test Results between Deontic and Nomenclature approaches for express mail

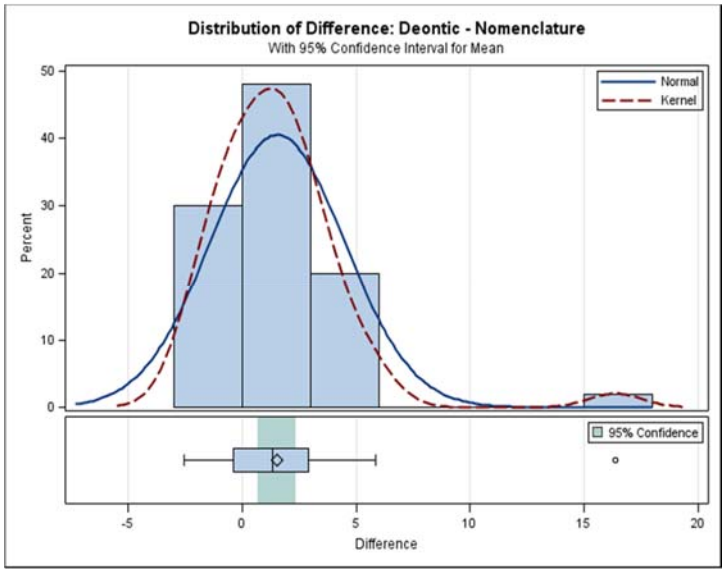


Figure 5: T-Test Results between Deontic and Nomenclature approaches for bulk parcel

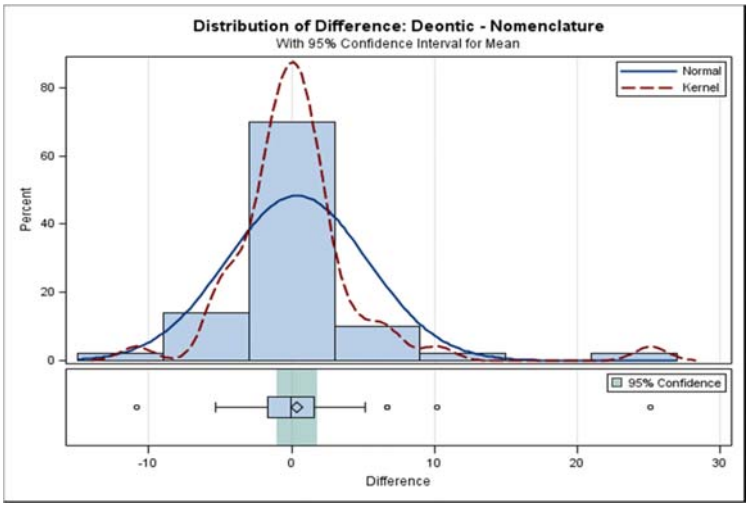


Figure 6: T-Test Results between Deontic and Nomenclature approaches for letter deliveries

From the results, four of six processes demonstrated that the new nomenclature approach is more efficient than the classical deontic approach. For the first case where nomenclature approach failed to outperform, the processes were in general very short and thus there are no significant amount of compression work.

Given the nature of the approach, the analysis can be scheduled to run for regular time intervals and the outputs compared to a predetermined list of possible processes for compliance checks. At the end of the daily operation, a report can be generated to calculate the compliancy rate.

Conclusion

The nomenclature approach offers improvement to the compliancy checking in the following area.

1. Innovative approach
2. Ease of logical interpretation
3. Detection of non-compliancy
4. Real time detection is possible
5. Improve productivity
6. Ensure compliance

The approach also allow flexibility to users who wants to further adapt it for their company usage. The ease of applying the compliancy check in operation also enables the management to have a tighter rein on compliancy as well as near real time report any compliancy failures. However, there are still some weaknesses in the existing approach. The

approach currently do not deal with multiple level nested process as these processes are extremely rare. Reseting the process or transfer from one process to another cannot be dealt with using this approach. Future research direction requires more in depth improvement to the heuristic to identify and address these issues.

References

A. Ghose and G. Koliadis. Auditing business process compliance. ICSOC, LNCS 4749, pages 169–180, 2007.

Basel III, Basel Committee on Banking Supervision (2011)

C. Abrams, J. von Kanel, S. Muller, B. Pfitzmann, and S. Ruschka-Taylor, ‘‘Optimized Enterprise Risk Management,’’ IBM Systems Journal 46, No. 2, 219–234 (2007).

C. Giblin, A. Y. Liu, S. Muller, B. Pfitzmann, and X. Zhou, ‘‘Regulations Expressed as Logical Models (REALM),’’ Proceedings of the 18th Annual Conference on Legal Knowledge and Information Systems, Brussels, Belgium (2005), pp. 37–48.

Control Objectives for Information and Related Technology (COBIT), Version 4.0 IT Governance Institute (2005), http://www.itgi.org.

F. Leymann and D. Roller, Production Workflow: Concepts and Techniques, Prentice Hall PTR, Upper Saddle River, NJ (2000).

G. Antoniou, D. Billington, G. Governatori, and M. J. Maher. Representation results for defeasible logic. ACM Trans. on Computational Logic, 2:255–287, 2001.

Governatori, Guido, and Antonino Rotolo. "A conceptually rich model of business process compliance." *Proceedings of the Seventh Asia-Pacific Conference on Conceptual Modelling-Volume 110*. Australian Computer Society, Inc., 2010.

Gramm-Leach-Bliley Act of 1999 (GLBA), Public Law 106- 102 (113 Statute 1338), United States Senate Committee on Banking, Housing, and Urban Affairs (1999).

IT Infrastructure Library (ITIL), Office of Government Commerce (2006), http://www.itil.co.uk.

Ilie, Lucian, and William F. Smyth. "Minimum unique substrings and maximum repeats." *Fundamenta Informaticae* 110.1 (2011): 183-195.

Ly, Linh Thao, et al. "Monitoring business process compliance using compliance rule graphs." *On the Move to Meaningful Internet Systems: OTM 2011*. Springer Berlin Heidelberg, 2011. 82-99.

M. Havey, Essential Business Process Modeling, O’Reilly & Associates, Sebastopol, CA (2005).

M. Alberti, D. Daolio, P. Torroni, M. Gavanelli, E. Lamma, and P. Mello. Specification and Verification of Agent Interaction Protocols in a Logic-based System. SAC’04, pages 72–78, 2004.

M. Alberti, M. Gavanelli, E. Lamma, P. Mello, P. Torroni, and G. Sartor. Mapping of Deontic Operators to Abductive Expectations. NORMAS, pages 126–136, 2005.

Sarbanes-Oxley Act of 2002, Public Law 107-204 (116 Statute 745), United States Senate and House of Representatives in Congress (2002).

The Money Laundering Regulations, Statutory Instrument 2003 No. 3075, Act of Parliament, http://www.opsi.gov.uk/si/si2003/20033075.htm.

Ukkonen, Esko. "On-line construction of suffix trees." *Algorithmica* 14.3 (1995): 249-260.

Biography

Murphy Choy is an instructor in Singapore Management University. He has extensive experience in the area of risk management and has developed credit risk models for several banks in various parts of the world. Murphy has also developed solutions and models for other industries such as logistics and retail. He is completing his doctorate and his research interest is in the area of Risk Management Operation management and Text Mining.

Nang Laik is a director of Master of IT in Business-Analytics (MITB-A) at School of Information Systems, Singapore Management University (SMU). She holds a PhD from Imperial College, London where her research focused on operations research (OR) in the area of optimization of resource. She teaches undergraduate core modules and master level courses in SMU. She has been teaching Business Process Modelling course and Computer as an Analysis Tool and Business Analytics Practicum. Her research expertise lies in the simulation and modelling of large scale real-world problems and the development of computationally efficient algorithms to enable sound and intelligent decision making in the organization. Nang Laik is an expert in transportation and logistics industry, she serves as a consultant for one of the best airports - Changi Airport Group to use data and decision Analytics to generate insights, make better decision and improve the business efficiency and productivity.

Koo Ping Shung is an experienced practitioner of analytics spanning numerous industries and business functional areas such as Marketing Management, Risk Management, Strategic Management and Human Resource Management. He specializes in the business adoption and implementation of analytics and is acknowledged by fellow practitioners as well-read in the field of business analytics. He is a trainer for SAS, a business analytic solution provider, in SAS-related

subject matters. He has excellent command of the SAS programming language with deep understanding of the functionality of the base SAS software. His career have taken him to two Singapore Universities, National Institute of Education and Singapore Management University and two banks in Singapore, DBS Bank and OCBC Bank.